

Newly Developed Structure-Based Methods Do Not Outperform Standard Sequence-Based Methods for Large-Scale Phylogenomics

Giacomo Mutti ^{1,2}, Eduard Ocaña-Pallarès ^{1,2,3,†,*}, Toni Gabaldón ^{1,2,4,5,†,*}

¹Barcelona Supercomputing Centre (BSC-CNS), Barcelona 08034, Spain

²Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona 08028, Spain

³Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Barcelona, Spain

⁴Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain

⁵CIBER de Enfermedades Infecciosas, Instituto de Salud Carlos III, Madrid, Spain

[†]Both authors share senior author contribution.

*Corresponding authors: E-mails: ed3716@gmail.com; toni.gabaldon@bsc.es.

Associate editor: Belinda Chang

Abstract

Recent developments in protein structure prediction have allowed the use of this previously limited source of information at genome-wide scales. It has been proposed that the use of structural information may offer advantages over sequences in phylogenetic reconstruction, due to their slower rate of evolution and direct correlation to function. Here, we examined how recently developed methods for structure-based homology search and tree reconstruction compare with current state-of-the-art sequence-based methods in reconstructing genome-wide collections of gene phylogenies (i.e. phylomes). While structure-based methods can be useful in specific scenarios, we found that their current performance does not justify using the newly developed structure-based methods as a default choice in large-scale phylogenetic studies. On the one hand, the best performing sequence-based tree reconstruction methods still outperform structure-based methods for this task. On the other hand, structure-based homology detection methods provide larger lists of candidate homologs, as previously reported. However, this comes at the expense of missing hits identified by sequence-based methods, as well as providing sets of homolog candidates with higher fractions of false positives. These insights help to guide the use of structural data in comparative genomics and highlight the need to continue improving structure-based approaches. Our pipeline is fully reproducible and has been implemented in a Snakemake workflow. This will facilitate a continuous assessment of future improvements of structure-based tools in the AlphaFold era.

Keywords: phylogenetics, structural phylogenetics, phylome, orthology, homology

Introduction

As the number of available genomes increases, so does our ability to investigate deeper and unresolved evolutionary questions. However, traditional phylogenetic methods based on sequences have limitations, such as the saturation of the phylogenetic signal when dealing with highly-divergent sequences and/or distantly related species. When seeking new sources of phylogenetic information, protein structures have been considered based on the idea that they are expected to diverge at a slower pace than sequences (Illergård et al. 2009). However, the scarcity of experimental structural information has limited their use in large-scale phylogenetic studies.

The development of AlphaFold2 has recently revolutionized the field of structural biology, unlocking the analyses of protein structures at scales previously considered unachievable (Jumper et al. 2021). As a consequence, new methods have been developed to handle such scales. One example is Foldseek, a software that allows fast alignment of protein structures based on a recoding of the structural information into a 20-state alphabet (Van Kempen et al. 2024). Recently, Moi et al. (2023) developed Foldtree: a pipeline to quickly

compute distance-based phylogenies from protein structures which the authors argued can outperform traditional methods also for closely related proteins.

Motivated by the potential shown by Foldtree, we analyzed whether a phylogenomic pipeline could be adapted to use protein structures for genome-wide phylogenetic analysis, such as the reconstruction of all evolutionary histories of human genes—i.e.—the human phylome (Huerta-Cepas et al. 2007).

A Pipeline to Benchmark Structure Versus Sequence Phylomes

A set of 18 eukaryotic species (supplementary table S1, Supplementary Material online) was selected to reconstruct the human phylome based on available proteomes from UniProt (The UniProt Consortium et al. 2023). By inferring the human phylome, we refer to reconstructing the phylogenetic relationships between all *H. sapiens* proteins (seed/query proteins) and their homologs in the species set (target proteins).

Protein structures were downloaded from the AlphaFold database (Varadi et al. 2024). Proteins with poorly predicted structures [average predicted Local Distance Difference Test

Received: December 20, 2024. Revised: April 1, 2025. Accepted: May 29, 2025

© The Author(s) 2025. Published by Oxford University Press on behalf of Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

(pLDDT) < 40] were excluded. Protein sequences were extracted from the structure files to ensure full congruence between sequences and structures (Fig. 1a).

The sequence and the structure of each *H. sapiens* seed protein were aligned with BlastP and Foldseek, respectively, against the 18 species' proteomes dataset (Altschul et al. 1990; Van Kempen et al. 2024). Both results were filtered based on E-value < 1e-3, query coverage > 50%, and target coverage > 50%. Four target sets were retrieved for each seed: top 150 BlastP hits (*Bp*), top 150 Foldseek hits (*Fs*), *Union*, and intersection (*Common*) between *Bp* and *Fs* sets (Fig. 1b).

For each target set, eight tree reconstruction methods were applied (Fig. 1c; [supplementary methods](#), [Supplementary Material](#) online for details):

1. *FT*: structure-based distance method. The same procedure as the Foldtree Snakemake implementation (Moi et al. 2023).
2. *3Di*: structure-based Maximum Likelihood (ML) method. Structures are recoded with the 3Di alphabet computed by Foldseek. After masking sites with pLDDT < 50, 3Di sequences were aligned with FoldMason (Gilchrist et al. 2024). The alignments were then trimmed with TrimAl -gappyout (Capella-Gutiérrez et al. 2009). Trees were reconstructed with IQ-TREE (Minh et al. 2020) specifying as substitution matrix the 3Di matrix computed in Puente-Lelievre et al. (2024).
3. *GTR*: ML structure-based method. The same alignment as *3Di* was used in IQ-TREE using the GTR substitution matrix (GTR20).
4. *AF*: ML structure-based method. The same alignment as *3Di* was used in IQ-TREE using the AF substitution matrix computed in (Garg and Hochberg 2025).
5. *LLM*: same as AF using the alternative substitution matrix computed in (Garg and Hochberg 2025).
6. *LG*: sequence-based ML method. Protein sequences were aligned with mafft -auto (Katoh and Standley 2013) and trimmed as in *3Di* and *GTR*. Trees were reconstructed with IQ-TREE specifying the LG substitution matrix (Le and Gascuel 2008).
7. *FM*: sequence-based distance method. The same alignment for *LG* was used as input to FastME (Lefort et al. 2015).
8. *Part*: the trimmed alignments from *LG* and *3Di* were concatenated and used as input in IQ-Tree using a partition scheme.

For computational reasons, we subsampled 1,000 random seeds, filtered those with < 4 *common* hits, and proceeded with inferring 32 phylogenies per seed (four target sets per eight tree reconstruction methods). We compared the accuracy of each gene tree reconstruction method according to three different metrics (see [supplementary methods](#), [Supplementary Material](#) online for details).

All the scripts used in this project were developed as a reproducible and customizable pipeline in Snakemake (Mölder et al. 2021) available at https://github.com/Gabaldonlab/structural_phylome. These analyses can be reproduced with any taxon set present in UniProt. Check the [supplementary methods](#), [Supplementary Material](#) online for additional information.

Divergences Between BlastP and Foldseek

Among the query-target pairs identified by BlastP and Foldseek, only 2.62×10^6 pairs (23.6%) were detected by both tools (*common* targets). After filtering, this proportion increases to 39.8% (0.58×10^6 pairs), highlighting a significant disparity between the two approaches (Fig. 1b). The two tools, respectively, identified 3.53×10^6 (31.8%) and 4.94×10^6 (44.5%) unique hits (referred to as *singletons* hereafter). These proportions expectedly decrease to 16.6% and 43.6% after filtering.

To understand these striking differences, we investigated three alignment metrics in the unfiltered results: percentage of identity, E-value and query coverage. For each metric we explored the correlation among BlastP and Foldseek in *common* hits (see binned scatterplots in Fig. 2a-c), and the distribution of values in *singletons* and *common* hits for these metrics (see marginal distributions in Fig. 2a-c). We observed that the percentage of identity of *common* hits is highly correlated between BlastP and Foldseek ($r = 0.85$, Fig. 2a). Interestingly, only Foldseek *singletons* have lower identity values (mode < 25%) when compared with *common* hits (Fig. 2a). This indicates that Foldseek, as proposed, may be useful to detect highly-divergent homologs in the so-called sequence identity “twilight zone” (Sander and Schneider 1991; Rost 1999; Puente-Lelievre et al. 2024). However, in an assessment based on the EggNOG orthologous groups (Huerta-Cepas et al. 2019), we found that the potential of Foldseek in identifying remote homologs comes at the expense of missing a higher fraction of established co-orthologous relationships when compared with BlastP ([supplementary fig. S1a](#), [Supplementary Material](#) online). The E-values of both tools show moderate correlation ($r = 0.43$) but Foldseek E-values tend to be lower (Fig. 2b). BlastP and Foldseek *singletons* have similar E-value distributions when compared with *common* hits. Regarding query coverage, BlastP *singletons* usually have lower values for this metric (see red marginal distribution in Fig. 2c). Consistent with this, the hits identified by BlastP but not by Foldseek in the EggNOG orthologous groups analysis usually show low query coverage values ([supplementary fig. S1b](#), [Supplementary Material](#) online). This suggests Foldseek may be missing hits when homology conservation is restricted to a smaller fraction of the protein space.

For each alignment, Foldseek computes both the average Local Difference Distance Test (LDDT) and the Template Modeling (TM) Score. These two metrics quantify the overlap between pairs of aligned structures and both are clearly lower in Foldseek *singletons* indicating highly-divergent homologs or possible spurious homologs (Fig. 2d, e). Regarding the quality of protein structure predictions, *common* hits have higher average pLDDT (as computed by AlphaFold) compared with both Foldseek and BlastP *singletons* (Fig. 2f). This indicates that poorly predicted structures may be responsible for both false positives (Foldseek *singletons*) and false negatives (BlastP *singletons*) in Foldseek searches.

As an orthogonal benchmark of the two similarity search methods, we explored the homogeneity of CATH annotations in filtered BlastP and Foldseek results. Our rationale was that correct homologous identifications should result in target sets with higher fractions of shared C, CA, CAT and CATH domain annotations with respect to the query. *Common* hits have more homogenous annotations, followed by BlastP *singletons* (Fig. 2g). The lower levels of concordance of Foldseek *singletons* suggest that this method recovers a higher fraction of non-homologous targets than BlastP.

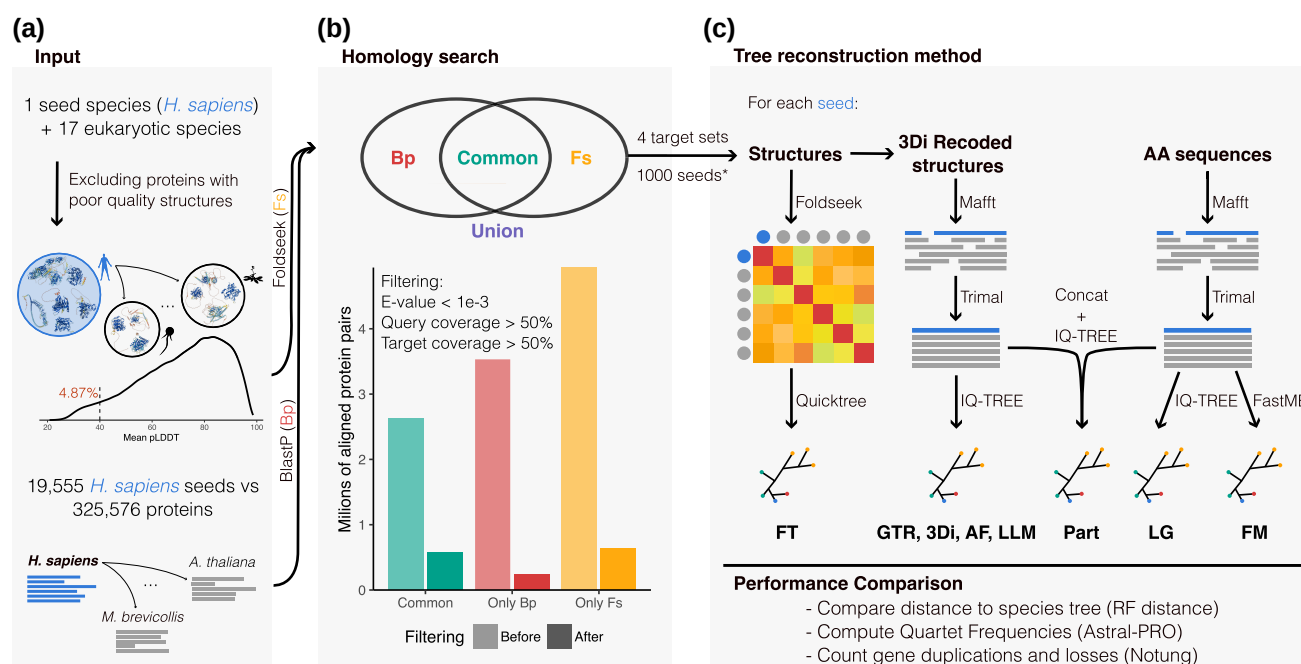


Fig. 1. Schematic representation of the pipeline. a) Primary amino acid sequences and 3Di-recoded structures from *Homo sapiens*' proteins (seeds) are aligned against a dataset of 18 eukaryotic species with BlastP (Bp) and Foldseek (Fs), respectively. b) Before entering into the phylogenetic pipeline, Bp and Fs results are divided into four target sets per seed as shown in the Venn diagram. The number of query-target pairs in each target set before and after filtering is shown below. c) The four target sets of each seed are submitted to eight tree reconstruction methods. (For computational reasons, we restricted step C to 1,000 randomly selected seeds). *Among the randomly selected seeds, only those with at least four common hits entered into step C.

Underperformance of Structure-Based Methods for Tree Reconstruction

We compared the performance of sequence-based versus structure-based tree reconstruction methods based on the following metrics (Fig. 3): (i) Robinson–Foulds (RF) (Robinson and Foulds 1981) distance to the species tree of single copy orthologs subtrees obtained from splitting multicopy gene families using DISCO; (ii) fraction of gene tree quartets agreeing with the species tree topology (quartet support), and (iii) number of duplication and losses needed to reconcile the gene trees with the species tree. More accurate gene trees are expected to have lower RF distances to the species tree, show higher quartet support values, and require less events to be reconciled.

Structure-based methods (3Di, GTR, FT, AF, and LLM) did not outperform sequence methods (LG, FM) in any of these metrics (Fig. 3). All the ML structure-based methods (3Di, GTR, AF, and LLM) clearly underperformed the ML sequence-based method (LG). Bootstrap support values were higher in LG, perhaps due to the worse performance of ML structure-based methods (supplementary fig. S2a, Supplementary Material online). Both distance methods, sequence-based (FM) and structure-based (FT), performed similarly (Fig. 3). Further, they also produced very similar trees to each other (supplementary fig. S2b, Supplementary Material online). This suggests structure and sequence data may provide similar phylogenetic information but that structural ML methods still have room for improvement. Despite the overall worse performance of ML structure-based methods, we confirm that the AF and LLM matrices provide a better fit than 3Di and GTR (as reported in Garg and Hochberg 2025), indicating that these recent contributions represent important steps toward a better ML structure-based inference (supplementary fig. S2c, Supplementary Material online).

LG outperformed structure-based and sequence-based distance methods, most clearly in metrics (i) and (ii) (Fig. 3a and b), inferring less discordant trees with the species tree for all target sets. This implies that LG performed better also for the target sets which include Foldseek *singletons* (i.e. candidate homologs only identified based on structural information). The better performance of ML sequence-based was confirmed in an independent benchmarking done with one of the datasets that were used to benchmark Foldtree (Moi et al. 2023) (supplementary fig. S3, Supplementary Material online, see supplementary methods, Supplementary Material online). Finally, we also explored a partition scheme combining sequence and 3Di information (*part*), as done in a recent study (Puentes-Lelievre et al. 2024). Interestingly, this condition outperformed ML structure-based methods but does not seem to improve the performance of the ML sequence-only condition (LG) despite being more computationally intensive and time consuming.

To complement the previous analyses done on the distinct target sets, we explored the topological information of the gene trees. Foldseek *singletons* show a greater normalized internode distance to the seed compared with BlastP *singletons*, a median distance of 0.625 and 0.547, respectively, based on LG, the best performing tree reconstruction method (this result is consistent in all tested conditions, supplementary fig. S2d, Supplementary Material online). *Common* hits were the closest to the seed (median = 0.455), showing a more similar distribution to BlastP *singletons* than to Foldseek *singletons* (supplementary fig. S2d, Supplementary Material online). Overall, these results are concordant with Foldseek being able to detect homologs in the “twilight zone”. However, this comes at the expense of identifying a higher fraction of targets with non-homologous structures than BlastP (Fig. 2g), and also of missing BlastP *singletons*, which

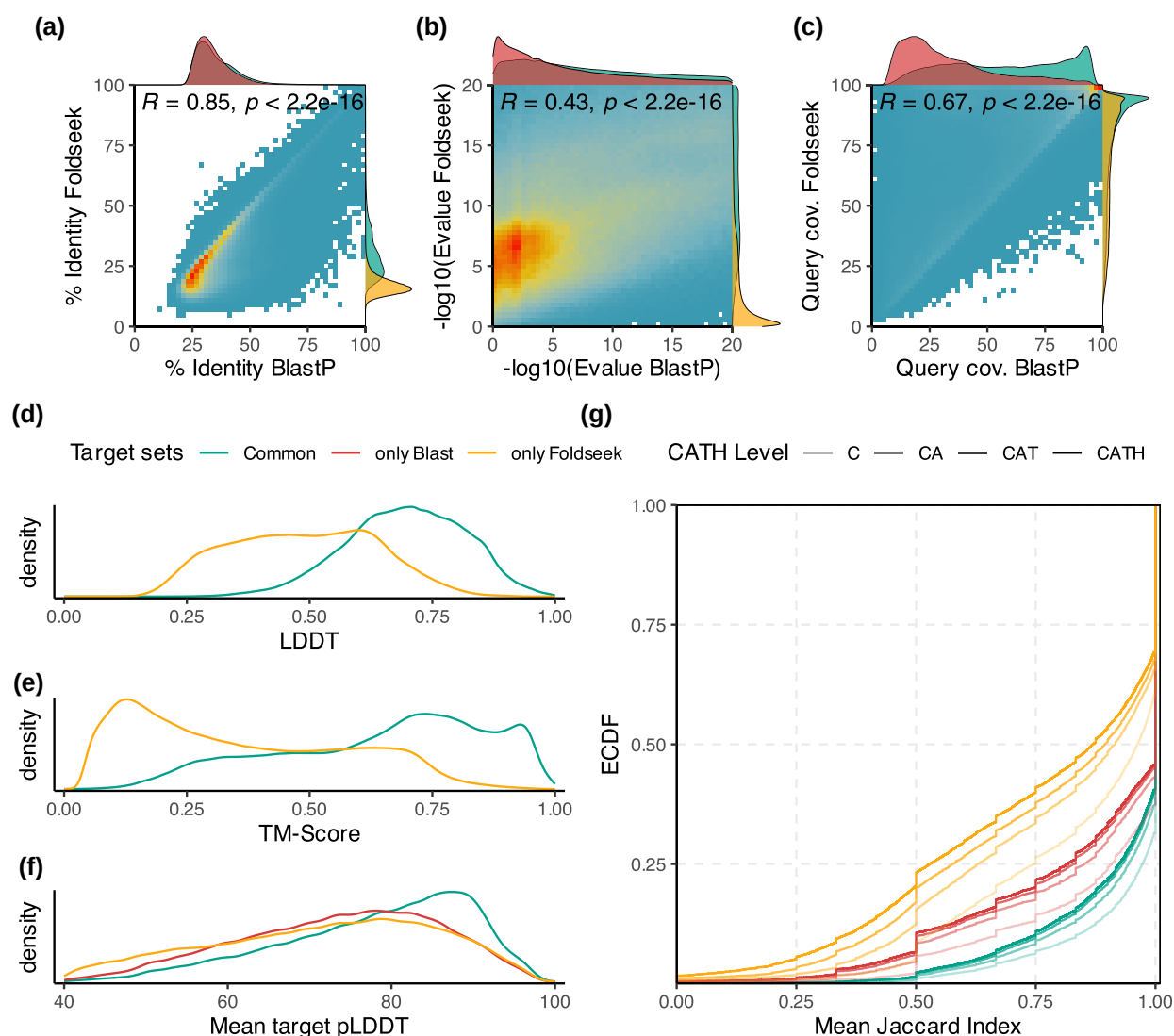


Fig. 2. Distributions and correlation of a) percentage identity, b) $-\log_{10}(\text{E-value})$ and c) query coverage for unfiltered BlastP and Foldseek results. The marginal distributions are color-coded according to the target set. Distribution of d) Local Difference Distance Test (LDDT), e) Template Modeling (TM) score and f) mean target predicted LDDT between different target sets. g) Cumulative distribution of mean Jaccard index per query for all levels of CATH annotation, including protein class (C), architecture (A), topology (T), and homologous superfamily (H) (indicated coded with different transparency levels, see legend). See [supplementary methods, Supplementary Material](#) online for details on this analysis.

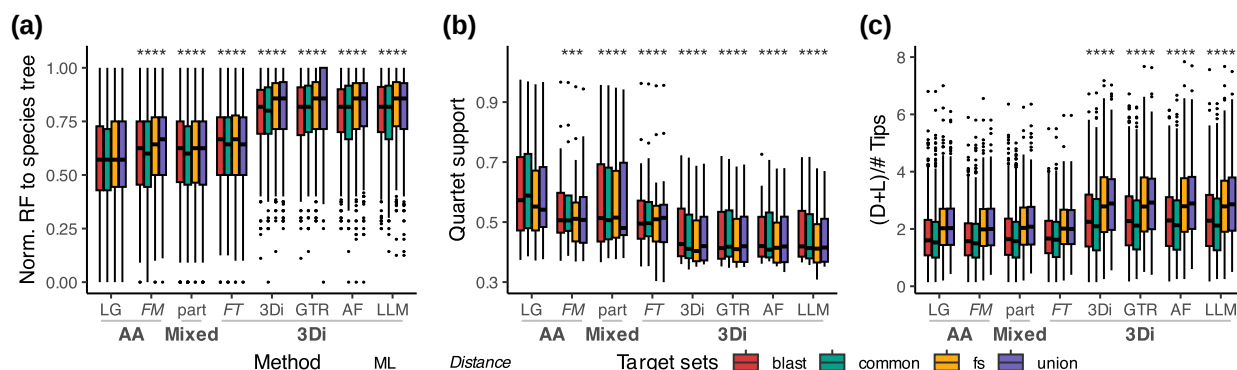


Fig. 3. Boxplot distribution grouped by target set and tree reconstruction method of a) normalised Robinson–Foulds (RF) distance of decomposed single copy gene trees to the species tree, b) First Quartet Frequency support values (this measure indicates how many times the nodes in the species trees are observed in the gene trees), and c) number of gene duplications and losses inferred by gene tree-species tree reconciliation normalised by number of tips. Significant differences to LG are annotated with asterisks over each respective boxplot. See the “Performance assessment of tree reconstruction methods” section of [supplementary Methods, Supplementary Material](#) online for details on how P -values were computed.

tend to be phylogenetically closer to the seed than Foldseek *singletons* (supplementary fig. S2d, Supplementary Material online).

As done in [Moi et al. \(2023\)](#), we assessed ultrametricity (measured as the variance of root-to-tip distances) as a possible index of tree reconstruction quality. We noticed that *FT* was considerably more ultrametric than any other method, as reported in [Moi et al. \(2023\)](#) (supplementary fig. S2e, Supplementary Material online). The four ML structure-based methods (*3Di*, *GTR*, *AF* and *LLM*) were more ultrametric than *LG*, possibly indicating that structures may, in fact, present a more “molecular clock-like” behavior than sequences. Notwithstanding this, since *LG* outperformed all structure-only based methods (Fig. 3), this indicates that ultrametricity is not necessarily an indicator of accurate phylogenetic reconstruction.

Conclusions

Before the development of AlphaFold, structural protein data was scarce, especially for non-model species (supplementary fig. S4, Supplementary Material online). Following the AlphaFold revolution, phylogenomic analyses based on structural information are now possible thanks to tools such as Foldseek and Foldtree. Despite this being a very important achievement, our results altogether point out that structure-based methods still do not outperform the standard sequence-based tools for large-scale phylogenetics, particularly in the phylogenetic reconstruction step. The potential of the structure-based tool Foldseek to detect highly-divergent homologs in the “twilight zone” (Fig. 2a) could be of particular interest when protein sequences provide limited resolution for homology search ([Himmel et al. 2023](#); [Köstlbacher et al. 2024](#)). However, its usage comes with the risk of missing sequences detected as homologs by Blast (BlastP *singletons*), which are more phylogenetically proximal to the seed sequence than Foldseek *singletons* (Fig. 2g; supplementary fig. S2d, Supplementary Material online). Regarding gene tree reconstruction, the ML sequence-based method performed better than *FT*, even when run on the set of structural homologs identified by Foldseek (Fig. 3). While ML sequence-based performed much better than ML structure-based for tree reconstruction, few differences in performance were observed between sequence and structure-based tree distance methods (Fig. 3). This suggests that future improvements either in the recoding of structural information ([Edgar 2024](#)) and/or in ML implementations of structure-based tree reconstruction methods ([Garg and Hochberg 2025](#); [Puentes-Lelievre et al. 2024](#)) could eventually lead structure-based methods to surpass sequence-based methods in large-scale phylogenomics.

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online.

Funding

G.M. received a predoctoral fellowship from the ‘la Caixa’ Foundation (LCF/BQ/DI22/11940014). E.O.-P. acknowledges support from FJC2021-046869-I funded by MCIN/AEI/10.13039/501100011033 and by “European Union NextGenerationEU/PRTR”, as well as from the Beatriz de Pinós programme (BP 2022, file number BP 00075). E.O.-P. also acknowledges that the project that gave rise to these results

received the support of a fellowship from the “la Caixa” Foundation (ID 100010434). The fellowship code is “LCF/BQ/PI24/12040009”. T.G. acknowledges support from the Spanish Ministry of Science and Innovation (grant numbers PID2021-126067NB-I00, CPP2021-008552, PCI2022-135066-2, and PDC2022-133266-I00), cofounded by ERDF “A way of making Europe”, as well as support from the Agència de Gestió d’Ajuts Universitaris i de Recerca (AGAUR) (grant number SGR01551), European Union’s Horizon 2020 research and innovation programme (grant number ERC-2016-724173); Gordon and Betty Moore Foundation (grant number GBMF9742); “la Caixa” foundation (grant number LCF/PR/HR21/00737), and Instituto de Salud Carlos III (IMPACT grant IMP/00019 and CIBERINFEC CB21/13/00061-ISCIII-SGEFI/ERDF).

Data Availability

All the data is already publicly available. The code used in this project, including how to recreate the same dataset and figures, is available at https://github.com/Gabaldonlab/structur al_phylome.

References

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215(3):403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2).
- The UniProt Consortium, Bateman A, Martin M-J, Orchard S, Magrane M, Ahmad S, Alpi E, Bowler-Barnett EH, Britto R, Bye-A-Jee H, et al. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Res.* 2023;51(D1):D523–D531. <https://doi.org/10.1093/nar/gkac1052>.
- Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. Trimal: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics.* 2009;25(15):1972–1973. <https://doi.org/10.1093/bioinformatics/btp348>.
- Edgar RC. Sequence alignment using large protein structure alphabets improves sensitivity to remote homologs. *bioRxiv* 2024.05.24.595840. <https://doi.org/10.1101/2024.05.24.595840>, 9 June 2024, preprint: not peer reviewed.
- Garg SG, Hochberg G. A General substitution matrix for structural phylogenetics. *Mol Biol Evol.* 2025;42(6). <https://doi.org/10.1093/molbev/msaf124>.
- Gilchrist CLM, Mirdita M, Steinegger M. Multiple protein structure alignment at scale with FoldMason. *bioRxiv* 2024.08.01.606130. <https://doi.org/10.1101/2024.08.01.606130>, 1 August 2024, preprint: not peer reviewed.
- Himmel NJ, Moi D, Benton R. Remote homolog detection places insect chemoreceptors in a cryptic protein superfamily spanning the tree of life. *Curr Biol.* 2023;33(22):5023–5033.e4. <https://doi.org/10.1016/j.cub.2023.10.008>.
- Huerta-Cepas J, Dopazo H, Dopazo J, Gabaldón T. The human phylome. *Genome Biol.* 2007;8(6):R109. <https://doi.org/10.1186/gb-2007-8-6-r109>.
- Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 2019;47(D1):D309–D314. <https://doi.org/10.1093/nar/gky1085>.
- Illergård K, Ardell DH, Elovsson A. Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. *Proteins Struct Funct Bioinf.* 2009;77(3):499–508. <https://doi.org/10.1002/prot.22458>.
- Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Židek A, Potapenko A, et al. Highly

- accurate protein structure prediction with AlphaFold. *Nature*. 2021;596(7873):583–589. <https://doi.org/10.1038/s41586-021-03819-2>.
- Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013;30(4):772–780. <https://doi.org/10.1093/molbev/mst010>.
- Köstlbacher S, Van Hooff JJE, Panagiotou K, Tamarit D, De Anda V, Appler KE, Baker BJ, Ettema TJG. Structure-based inference of eukaryotic complexity in *Asgard archaea*. *bioRxiv* 2024. 07.03.601958. <https://doi.org/10.1101/2024.07.03.601958>, 5 July 2024, preprint: not peer reviewed.
- Le SQ, Gascuel O. An improved general amino acid replacement matrix. *Mol Biol Evol*. 2008;25(7):1307–1320. <https://doi.org/10.1093/molbev/msn067>.
- Lefort V, Desper R, Gascuel O. FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. *Mol Biol Evol*. 2015;32(10):2798–2800. <https://doi.org/10.1093/molbev/msv150>.
- Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, Lanfear R. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol*. 2020;37(5):1530–1534. <https://doi.org/10.1093/molbev/msaa015>.
- Moi D, Bernard C, Steinegger M, Nevers Y, Langleib M, Dessimoz C. Structural phylogenetics unravels the evolutionary diversification of communication systems in gram-positive bacteria and their viruses. *bioRxiv* 2023.09.19.558401. <https://doi.org/10.1101/2023.09.19.558401>, 23 September 2023, preprint: not peer reviewed.
- Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, Forster J, Lee S, Twardziok SO, Kanitz A, *et al*. Sustainable data analysis with Snakemake. *F1000Res*. 2021;10:33. <https://doi.org/10.12688/f1000research.29032.2>.
- Puente-Lelievre C, Malik AJ, Douglas J, Ascher D, Baker M, Allison J, Poole A, Lundin D, Fullmer M, Bouckert R, *et al*. Tertiary-interaction characters enable fast, model-based structural phylogenetics beyond the twilight zone. *bioRxiv* 2023.12.12.571181. <https://doi.org/10.1101/2023.12.12.571181>, 9 January 2024, preprint: not peer reviewed.
- Robinson DF, Foulds LR. Comparison of phylogenetic trees. *Math Biosci*. 1981;53(1-2):131–147. [https://doi.org/10.1016/0025-5564\(81\)90043-2](https://doi.org/10.1016/0025-5564(81)90043-2).
- Rost B. Twilight zone of protein sequence alignments. *Protein Eng Des Select*. 1999;12(2):85–94. <https://doi.org/10.1093/protein/12.2.85>.
- Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins Struct Funct Bioinf*. 1991;9(1):56–68. <https://doi.org/10.1002/prot.340090107>.
- Van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, Söding J, Steinegger M. Fast and accurate protein structure search with Foldseek. *Nat Biotechnol*. 2024;42(2):243–246. <https://doi.org/10.1038/s41587-023-01773-0>.
- Varadi M, Bertoni D, Magana P, Paramval U, Pidruchna I, Radhakrishnan M, Tsenkov M, Nair S, Mirdita M, Yeo J, *et al*. AlphaFold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res*. 2024;52(D1):D368–D375. <https://doi.org/10.1093/nar/gkad1011>.