Newly developed structure-based methods do not outperform standard sequence-based methods for large-scale phylogenomics

Giacomo Mutti[1,2], Eduard Ocaña-Pallarés[1,2,*], and Toni Gabaldón[1,2,3,4*]

1) Barcelona Supercomputing Centre (BSC-CNS). Plaça Eusebi Güell, 1-3, 08034 Barcelona, Spain.

2) Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac, 10, 08028 Barcelona, Spain

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

4) CIBER de Enfermedades Infecciosas, Instituto de Salud Carlos III, Madrid, Spain

*Both authors share correspondence: eduard.ocana@bsc.es, toni.gabaldon@bsc.es

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 to current state-of-the-art sequence-based methods in reconstructing genome-wide collections of gene phylogenies (i.e. phylomes). Our findings reveal that sequence-based methods outperform current structure-based methods both in homology search and accuracy of phylogenetic reconstruction. While structure-based methods can be useful in specific scenarios, their current performance does not justify transitioning from sequence to structured-based methods in large-scale phylogenetic studies. These insights help guide the use of structural data in comparative genomics and highlight the need to continue improving structure-based approaches.

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 for 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 structural information has limited their use in large-scale phylogenetic studies.

The development of AlphaFold2 has recently revolutionised 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-states alphabet (van Kempen et al. 2023). More 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 analysed 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 (Tab. S1) was selected to reconstruct the human phylome based on available proteomes from UniProt (The UniProt Consortium et al. 2022). 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. 2023). Proteins with poorly predicted structures (average predicted Local Difference Distance Test (pLDDT) < 40) were excluded. Protein sequences were extracted from the structure files to ensure full congruence between sequences and structures (Fig. 1A).

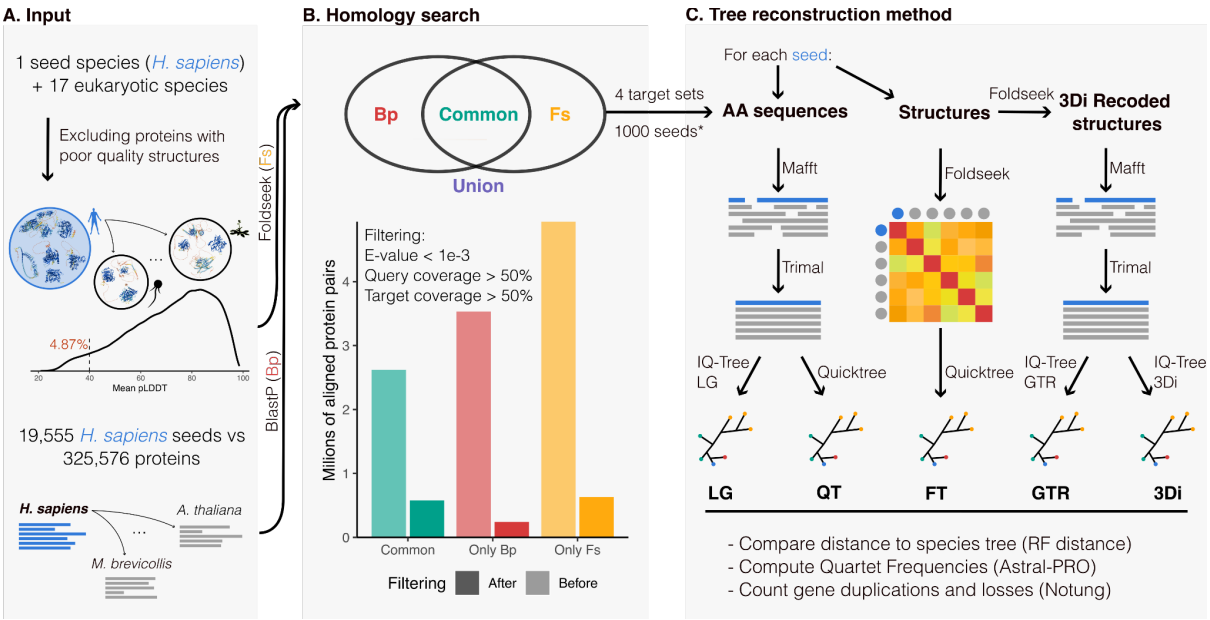

Figure 1. Schematic representation of the pipeline. A) Homologous sequences and structures to *Homo sapiens*' proteins (seed sequences) are aligned against a dataset of 18 eukaryotic species. B) Before entering into the phylogenetic pipeline, BlastP and Foldseek results are divided into four target sets per seed as shown in the Venn diagram. The number of query-targets pairs in each target set before and after filtering is shown below. C) The four target sets are submitted to five tree reconstruction methods. (For computational reasons, we restricted step C to 1000 randomly selected seeds. *Among the randomly selected seeds, only those with at least 4 common hits entered into step C.

The sequence and the structure of each *H. sapiens* seed protein was aligned with BlastP and Foldseek, respectively, against the 18 species' proteomes database (Altschul et al. 1990; van Kempen et al. 2023). 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 (*blast*), top 150 foldseek hits (*fs*), *union*, and intersection (*common*) between *blast* and *fs* sets (Fig. 1B). We measured the consistency of both CATH ('Homologous superfamily class', as inferred by Gene3D) and Pfam domain annotations of the retrieved hits with respect to the seed as a proxy of the fraction of real homologs represented in each target set (Lewis et al. 2018; Mistry et al. 2020).

For each target set, five tree reconstruction methods were applied (Fig. 1C):

1. *FT*: structure-based distance method. The same procedure as 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 mafft specifying the 3Di scoring matrix used by Foldseek. The alignments were then trimmed with TrimAl -gappyout (Capella-Gutiérrez et al. 2009). Trees were reconstructed with IQ-Tree2 (Minh et al. 2020) specifying as substitution matrix the 3Di matrix computed in Puente-Lelievre et al. 2024.
3. *GTR*: structure-based ML method. The same alignment as *3Di* was used in IQ-Tree2 using the GTR substitution matrix (GTR20).
4. *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-Tree2 specifying the LG substitution matrix (Le and Gascuel 2008).

5. *QT*: sequence-based distance method. The same alignment for *LG* was used as input to QuickTree (Howe et al. 2002).

For computational reasons we subsampled 1000 random seeds, filtered those with less than 4 *common* hits, and proceeded with inferring 20 phylogenies per seed (four target sets per five tree reconstruction methods). We compared the accuracy of sequence-based vs structure-based gene tree reconstruction methods by measuring the degree of discordance of the resulting gene trees with respect to the species tree (see Supplementary methods for details).

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

Divergences between BlastP and Foldseek

Among the query-target pairs found by BlastP and Foldseek, only 23.6% (39.8% after filtering) were found by both tools (*common* targets), indicating a large disparity between the two approaches (Fig. 1B). The two tools respectively found 31.8% and 44.5% (16.6% and 43.6% after filtering) of unique hits, referred as *singletons* hereafter.

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 to *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, this comes at the expense of missing hits with higher identity values that are only detected by BlastP (Fig. 2A). 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 to *common* hits. Regarding query coverage, BlastP singletons generally cover less than 50% of the query (Fig. 2C). This suggests Foldseek may be missing hits when homology conservation is restricted to a smaller fraction of the protein space.

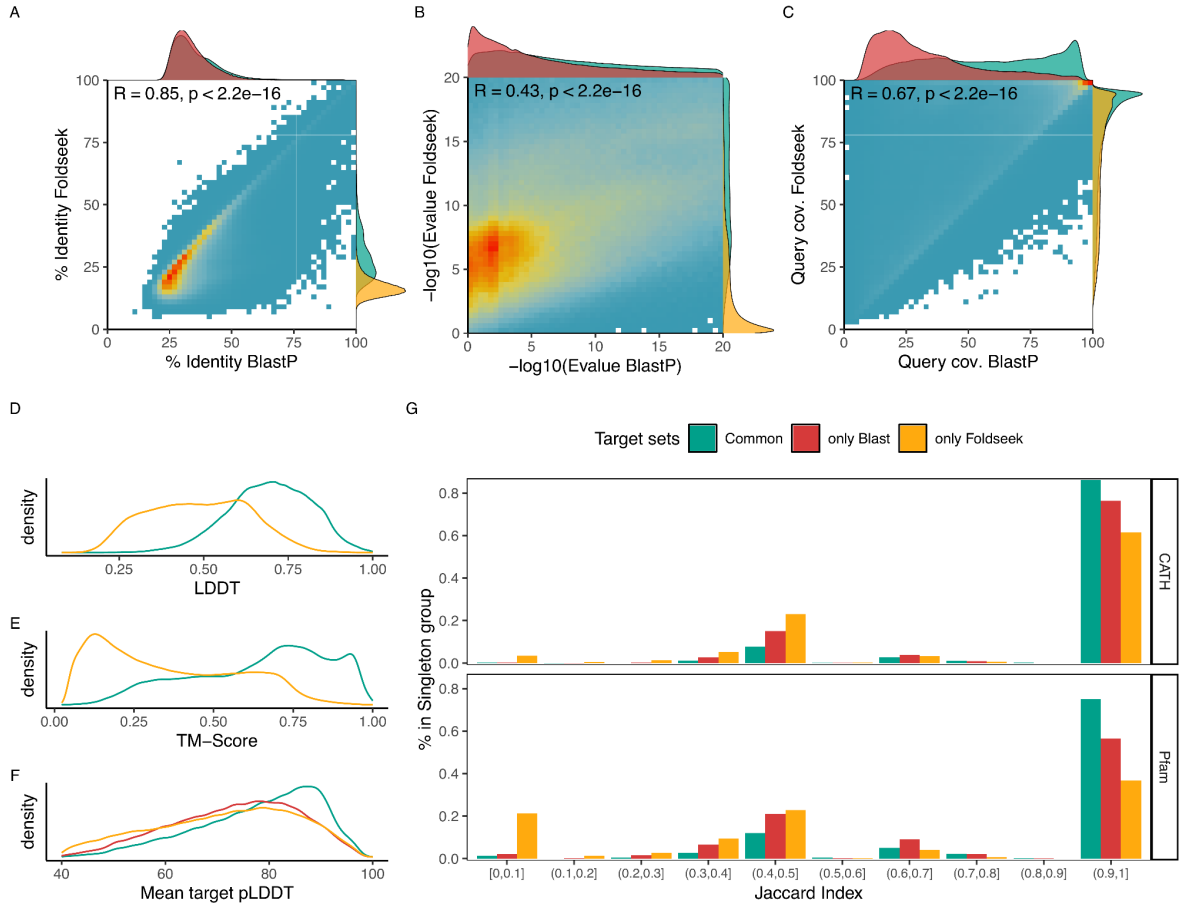

Figure 2. Distributions and correlation of A) percentage identity, B) -Log10E-value and C) query coverage for unfiltered BlastP (x-axis) and Foldseek (y-axis) results. The binned scatter plot colour scale goes from light blue (low density) to red (high density). The marginal distributions are colour-coded according to the target set (see colour legend on the top of panel G). Distribution of D) Local Difference Distance Test (LDDT), E) Template-Modelling score (TM-score) and F) mean target predicted LDDT between different target sets of Foldseek results. G) Histogram distribution of Jaccard index values assessing consistency on CATH and Pfam annotations between seeds and target sequences, binned and coloured by target set.

For each alignment Foldseek computes both the average Local Difference Distance Test (LDDT) and the Template Modelling Score (TM). 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 to 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 and Pfam domain annotations in filtered BlastP and Foldseek results. Our rationale was that correct homologous identifications should result in shared CATH or Pfam domains for the hits of a given query. To account for multi-domain proteins we computed the Jaccard index of the domains shared by query and target (1 if both proteins share the same domain architecture). *Common* hits had a Jaccard higher than 0.9 in more than 80% of the cases while the value drops in both *singleton* classes but particularly in the Foldseek ones (Fig. 2G). Similar results were obtained with CATH (Fig. 2G). The lower levels

of concordance of Foldseek *singletons* suggests that this method recovers a higher fraction of non-homologous targets than BlastP.

Underperformance of structure-based methods for tree reconstruction

We compared the performance of sequence-based vs structure-based tree reconstruction methods based on the following metrics (Fig. 3): (1) RF distance to the species tree of one to one ortholog trees decomposed from all gene trees; (2) fraction of gene tree quartets agreeing with the species tree topology and (3) 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 values to the species tree, present higher quartet support values, and require less events to be reconciled.

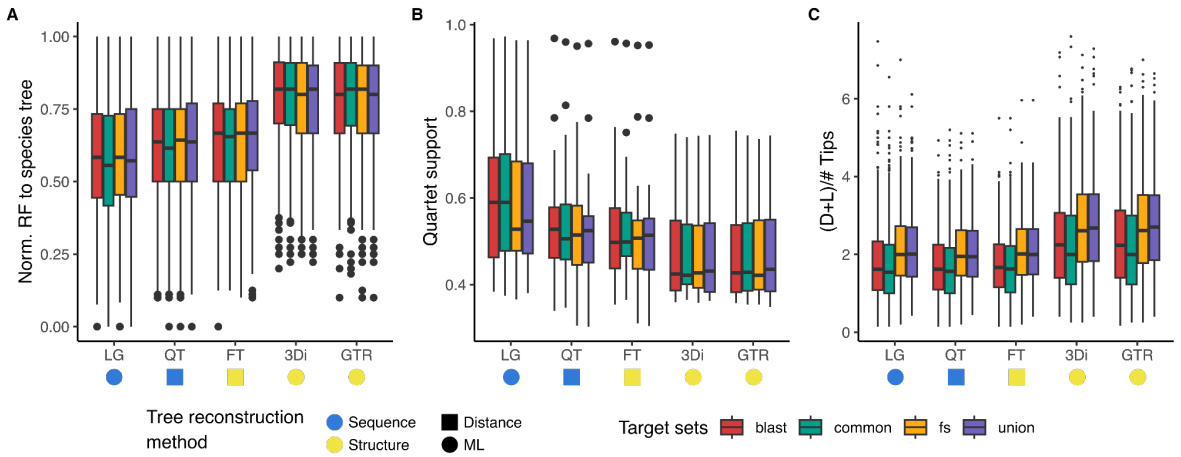

Figure 3. Distribution grouped by target set and tree reconstruction method of A) normalised Robinson-Foulds (RF) distance of decomposed single copy gene trees to the species tree, B) First Quartet Frequency support values (this measure indicates how many times the nodes in the species trees are observed in the gene trees, see Fig. S2 for more details) and C) number of gene duplications and losses inferred by gene tree-species tree reconciliation normalised by number of tips.

Structure-based methods (*3DI*, *GTR*, *FT*) did not outperform sequence methods (*LG*, *QT*) in any of these metrics (Fig. 3). The ML structure-based methods (*3DI*, *GTR*) clearly underperformed the ML sequence-based method (*LG*). In contrast, both distance methods, sequence-based (*QT*) and structure-based (*FT*), performed very similarly (Fig. 3). They also produced the most similar trees to each other (Fig. S1B). This suggests that structure and sequence may provide similar phylogenetic information but that structural ML methods still have room for improvement. Interestingly, bootstrap support values were higher in *LG* compared to *3Di* and *GTR* (Fig. S1A). This may relate to the worse performance of both ML structure-based approaches.

LG outperformed both distance methods in (1) and (2), inferring less discordant trees with the species tree for all target sets (Fig. 3A). 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). *LG* was also the best method in 9 out of the 15 internal nodes of the species tree, whereas *QT* was the best in the other six, with very similar values to *FT* (Fig. 3B, Fig. S2).

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 normalised

internode distance to the seed compared to BlastP *singletons*, median distance of 0.74 and 0.69, respectively, based on *LG*, the best performing tree reconstruction method (this result is consistent in all tested conditions, Fig. S1C). *Common* hits were the closest to the seed (median=0.60), showing a more similar distribution to BlastP *singletons* than to Foldseek *singletons* (Fig. S1C). 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 tend to be phylogenetically closer to the seed than Foldseek *singletons* (Fig. S1C).

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 both distance methods (*FT* and *QT*) were considerably more ultrametric than the ML methods. This implies that previously reported differences in ultrametricity between *FT* and sequence-based methods (Moi et al. 2023) are largely explained by a tendency of distance-based methods to recover more ultrametric trees (Fig. S1D). Still, the two ML structure-based methods (*3Di* and *GTR*) were more ultrametric than *LG*, possibly indicating that structures may in fact present a more "molecular clock-like" behaviour than sequences. Notwithstanding this, *LG* was the best method (Fig. 3), indicating that ultrametricity is not necessarily an indicator of accurate phylogenetic reconstruction.

Conclusions

Altogether, our results indicate that ML sequence-based methods outperform currently available structure-based methods in large-scale phylogenetic analyses. The potential of 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 a substantial fraction of sequences detected as homologs by Blast (BlastP *singletons*) which are more phylogenetically proximal to the seed sequence than Foldseek *singletons* (Figs. 2G and S1C). 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, little differences in performance were observed between sequence and structure-based tree distance methods (Fig. 3). This suggests that future improvements either in the recodification of structural-information (Edgar 2024) and/or in ML implementations of structure-based tree reconstruction methods (Puente-Lelievre et al. 2024) could eventually lead structure-based methods to surpass sequence-based methods in large-scale phylogenomics.

Funding

G.M. received a predoctoral fellowship from the "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 *Beatriu de Pinós* programme (BP 2022, file number BP 00075). **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 Catalan Research Agency (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 and Code Availability

All the code used in this project, including how to recreate the same dataset and figures, is available at https://github.com/Gabaldonlab/structural_phylome.

Supplementary Figures

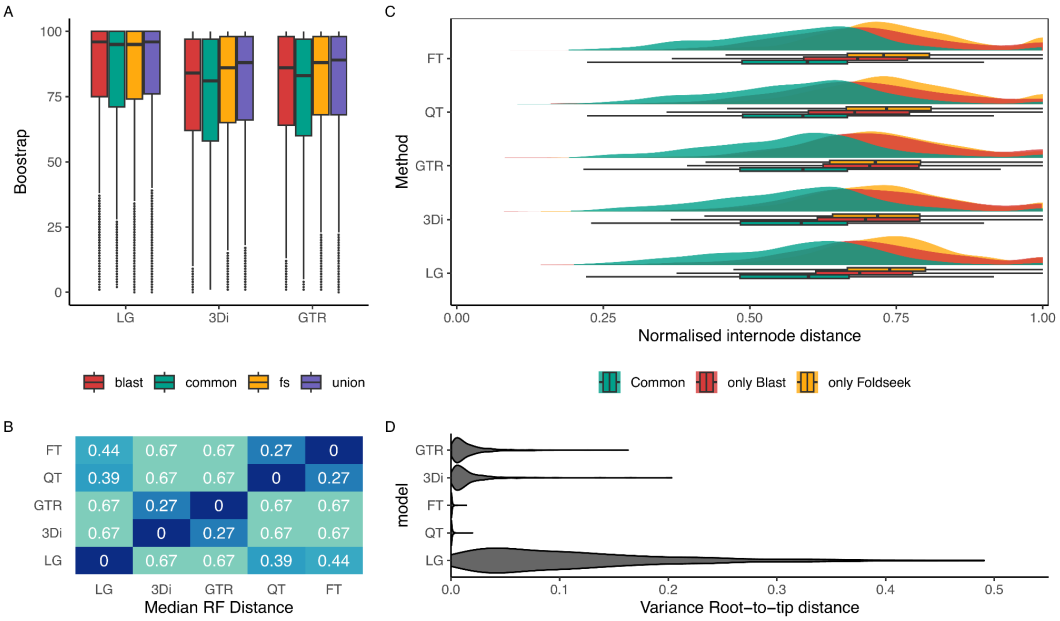

Figure S1. A) IQ-Tree ultrafast bootstrap values distribution binned by target set. B) Differences between tree reconstruction methods as measured by normalised Robinson-Foulds (RF) (median RF distance values are shown in the heatmap). C) Normalised internode distance to the seed (number of nodes from tip to seed divided by longest possible path to tip) distribution, coloured by target set and binned by tree reconstruction method. For the target sets of this panel, 'only Blast' and 'only Foldseek' refer to BlastP *singletons* and to Foldseek *singletons*, respectively. D) Violin plot distribution of root-to-tip distances (in midpoint rooted trees) for common hits. Each violin is scaled in order to have the same maximum width.

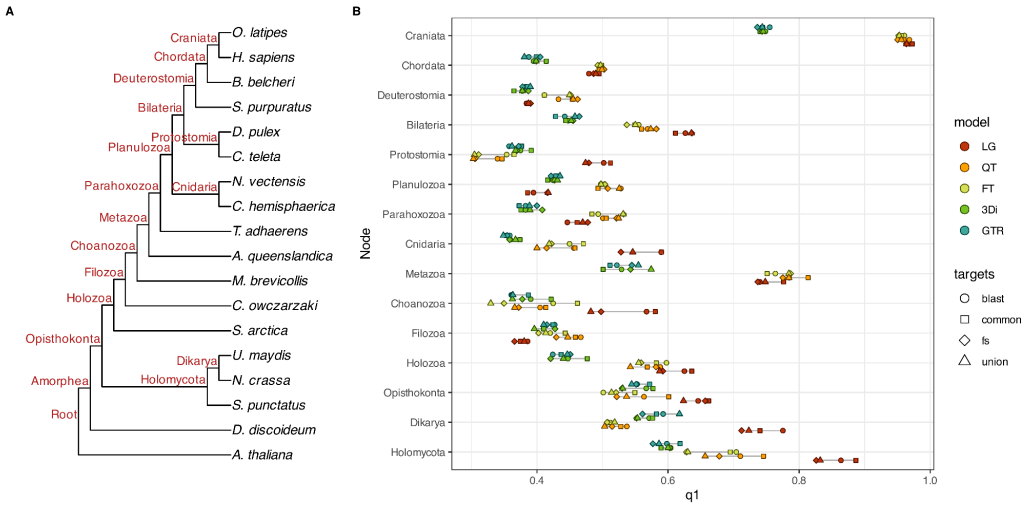

Figure S2. A) Species tree representation of the 18 eukaryotic species included in the dataset. See Supplementary Table 1 for more details. B) First Quartet Frequency for each node in the species tree as computed by Astral-Pro, grouped by target set and by tree reconstruction method. This measure indicates how many times the topology given the quartets of the nodes in the species tree is observed in the gene trees, normalised by the total number of trees where the species of the quartet are observed. The higher this number, the more concordant gene trees are with the species tree.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J. Mol. Biol.* [Internet] 215. Available from: https://pubmed.ncbi.nlm.nih.gov/2231712/

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25:1972–1973.

Edgar RC. 2024. Sequence alignment using large protein structure alphabets improves sensitivity to remote homologs. *bioRxiv* [Internet]:2024.05.24.595840. Available from: https://www.biorxiv.org/content/10.1101/2024.05.24.595840v2.abstract

Himmel NJ, Moi D, Benton R. 2023. Remote homolog detection places insect chemoreceptors in a cryptic protein superfamily spanning the tree of life. *Curr. Biol.* 33:5023–5033.e4.

Howe K, Bateman A, Durbin R. 2002. QuickTree: building huge Neighbour-Joining trees of protein sequences. *Bioinformatics* 18:1546–1547.

Huerta-Cepas J, Dopazo H, Dopazo J, Gabaldón T. 2007. The human phylome. *Genome Biol.* 8:R109.

Illergård K, Ardell DH, Elofsson A. 2009. Structure is three to ten times more conserved than sequence--a study of structural response in protein cores. *Proteins* 77:499–508.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596:583–589.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30:772–780.

van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, Söding J, Steinegger M. 2023. Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* 42:243–246.

Köstlbacher S, van Hooff JJE, Panagiotou K, Tamarit D, De Anda V, Appler KE, Baker BJ, Ettema TJG. 2024. Structure-based inference of eukaryotic complexity in Asgard archaea. *bioRxiv* [Internet]:2024.07.03.601958. Available from: https://www.biorxiv.org/content/10.1101/2024.07.03.601958v1.abstract

Le SQ, Gascuel O. 2008. An Improved General Amino Acid Replacement Matrix. *Mol. Biol. Evol.* 25:1307–1320.

Lewis TE, Sillitoe I, Dawson N, Lam SD, Clarke T, Lee D, Orengo C, Lees J. 2018. Gene3D: Extensive prediction of globular domains in proteins. *Nucleic Acids Res.* 46:D1282.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* 37:1530–1534.

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, et al. 2020. Pfam: The protein families database in 2021. *Nucleic Acids Res.* 49:D412–D419.

Moi D, Bernard C, Steinegger M, Nevers Y, Langleib M, Dessimoz C. 2023. Structural phylogenetics unravels the evolutionary diversification of communication systems in gram-positive bacteria and their viruses. *bioRxiv* [Internet]:2023.09.19.558401. Available from: https://www.biorxiv.org/content/10.1101/2023.09.19.558401v3.abstract

Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, Forster J, Lee S, Twardziok SO, Kanitz A, et al. 2021. Sustainable data analysis with Snakemake. *F1000Res.* 10:33.

Puente-Lelievre C, Malik AJ, Douglas J, Ascher D, Baker M, Allison J, Poole A, Lundin D, Fullmer M, Bouckert R, et al. 2024. Tertiary-interaction characters enable fast, model-based structural phylogenetics beyond the twilight zone. *bioRxiv* [Internet]:2023.12.12.571181. Available from: https://www.biorxiv.org/content/10.1101/2023.12.12.571181v2.abstract

Rost B. 1999. Twilight zone of protein sequence alignments. *Protein Eng.* 12:85–94.

Sander C, Schneider R. 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* 9:56–68.

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. 2022. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* 51:D523–D531.

Varadi M, Bertoni D, Magana P, Paramval U, Pidruchna I, Radhakrishnan M, Tsenkov M, Nair S, Mirdita M, Yeo J, et al. 2023. AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res.* 52:D368–D375.