Supplementary text to the paper “PhyloBench: a benchmark for evaluating phylogenetic programs”

SERGEY SPIRIN*, ANDREY SIGORSKIKH, ALEKSEY EFREMOV, DMITRY PENZAR AND ANNA KARYAGINA

* *Corresponding author, sas@belozersky.msu.ru*

MATERIALS AND METHODS: ADDITIONAL DETAILS

Algorithm for forming species sets

We created 12 sets of 60 species each using three different procedures. In the following description, the word “species” refers to an organism (either species or strain) having its own Uniprot mnemonic identification code, see https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/docs/speclist . The word “domain” denotes a Pfam [1] evolutionary domain in a particular protein. Each domain belongs to some protein of some species and has coordinates (start and end) in the sequence of this protein. Domains are organized into families according to Pfam database. We used information on domains, their families and their coordinates from the file “swisspfam” (http://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam33.1/swisspfam.gz). This file contains information on domains annotated in proteins from Uniprot reference proteomes.

Procedure 1. — To construct each set:

1. An embracing taxon was selected, i.e., Archaea or Fungi etc.
2. A taxonomic level, lower than the level of the embracing taxon, was chosen, i.e., “species” or “genus” or “family” etc.
3. From the embracing taxon, 60 species were selected under the following conditions:
   (a) The species belong to different taxa of the chosen level. For example, if at stage 2 the level of genus was chosen, all selected species should be from different geni.

(b) The number of common Pfam families, i.e., families represented by domains in all selected species, is maximum possible, among all choices of species from the embracing taxon satisfying the condition (a).

So, each species set of the benchmark is characterized by an embracing taxon and a taxonomic level. The choice of species (stage 3) is an NP-hard problem; thus, we applied a heuristic algorithm that provides not an absolute maximum of common Pfam families but a good approximation to the maximum.

Algorithm for selecting species. — Given an embracing taxon T, a taxonomic level L and a number of species *N* (for the Procedure 1 *N* equals 60), the following procedure was applied.

1. Select all subtaxa of T of the level L that contain species with proteins presented in "swisspfam".
2. Create a graph with weighted edges. The vertices of the graph are subtaxa selected on the previous step. Each two vertices are connected with an edge. The weight of an edge connecting A and B is the number of Pfam families represented in both subtaxa A and B simultaneously.
3. Repeat while the number of vertices is greater than N:
   (a) Choose vertex V_{min} with the minimum sum of weights of adjacent edges;
   (b) Remove V_{min} from the graph.
4. From each of the remaining vertices (i.e., *N* subtaxa of level L) choose a species with the maximum number of Pfam families annotated in its proteins.
5. Compose the list of common Pfam families for the selected species.

The steps 2 and 3 were performed three times in three variants, The first variant is literally the same as described; for the second variant the weights on the step 2 are numbers of Pfam families represented in A and B and having representatives in more than *N*/2 selected subtaxa; for the third variant the weights on the step 2 are numbers of Pfam families represented in A and B and having representatives in at least *N* selected subtaxa. The result of the entire procedure is the result of the variant that provides the largest number of common Pfam families.

Seven species sets created with Procedure 1 are listed in Table S1. Each species set corresponds to one set of orthologous groups (OGs) and to three alignment sets: with 15, 30, and 45 sequences in each alignment.

Table S1. Species sets made with Procedure 1

ID	Embracing taxon	Level	Number of Pfam families	Number of orthologous groups
AR	Archaea	species	349	363
AG	Archaea	genus	280	286
FI	Firmicutes	genus	464	490
AC	Actinobacteria	genus	728	833
AS	Ascomycota	genus	2205	3026
CH	Chordata	family	3584	3269
ST	Streptophyta	genus	2741	4172

We also constructed five other species sets by two modifications of the described procedure. Due to uneven distribution of sequenced species in taxa, the Procedure 1 sometimes produces highly unbalanced species sets. For example, applying this procedure to Metazoa leads to prevalence of Chordata in the result, and applying it to Fungi leads to prevalence of Ascomycota. Procedures 2 and 3 were used for “balancing” the selection of species of Eukaryota (species set EB), Metazoa (MA), Fungi (FB) and Proteobacteria (PB). Also, the species set OB (for “Other Bacteria”) was created to cover some bacterial phyla other than the most represented Actinobacteria, Firmicutes, and Proteobacteria. Currently any other bacterial phylum does not provide enough species represented in Pfam, thus we made one more species set of three phyla, Bacteroidetes, Cyanobacteria, and Acidobacteria. together.

Procedure 2. — Several taxons that are not subsets of each other (e.g., Metazoa and Fungi) were selected. For each selected taxon T, a number n_T was chosen, such that the sum of these numbers was 60. Also for each selected taxon a taxonomic level (lower than its own level) was chosen. Then Procedure 1 was performed for the first taxon as embracing, with the difference that we now selected not 60, but n_1 species. After this the Procedure 1 was performed for the second

taxon selecting n_2 species and with the difference that at the stage 3 we maximized the number of Pfam families that are presented in all species of first and second sets together. Similarly, we proceeded, at each step maximizing the number of Pfam families presented together in all previously selected species and in species selected from the current taxon. With this procedure we created two species sets, PB and OB, see Table S2.

Procedure 3. — This procedure is the same as the Procedure 2, but at some steps instead of an entire taxon we used a taxon with some subtaxa excluded. In this way, we created three species sets, MA, FB, and EB, see Table S2.

Table S2. Species sets made with Procedures 2 and 3

ID	Taxons	Level	Number of species	Number of Pfam33 families	Number of orthologous groups
PB	Epsilonproteobacteria	species	12	599	596
	Deltaproteobacteria	genus	12		
	Alphaproteobacteria	genus	12		
	Betaproteobacteria	genus	12		
	Gammaproteobacteria	family	12		
OB	Acidobacteria	species	5	580	603
	Bacteroidetes	genus	35		
	Cyanobacteria	genus	20		
MA	Mollusca	family	7	1194	959
	Arthropoda	order	12		
	Nematoda	family	12		
	Metazoa without {Arthropoda, Chordata, Mollusca, Nematoda}	genus	15		
	Mammalia	order	2		
	Aves	order	2		
	Actinopteri	order	2		
	Chordata without {Mammalia, Aves, Actinopteri}	order	8		
FB	Ascomycota	genus	26	1722	1993
	Basidiomycota	genus	26		
	Fungi without Ascomycota and Basidiomycota	genus	8		
EB	Metazoa	phylum	12	909	856
	Streptophyta	order	12		
	Fungi	class	12		
	Eukaryota without {Metazoa, Streptophyta, Fungi}	genus	24		

Algorithm for building species trees

The species tree for each species set was obtained as follows.

1. Sequences of each orthologous group were aligned by MUSCLE [2] with defaults.
2. If the alignment contains 10 or more columns that are not 100% conserved and do not contain gaps, three trees were inferred: (i) By the program TNT [3] using command "*mult*" with default parameters; (ii) By the program RAxML 8.2.12 [4] with the option –m PROTGAMMAAUTO, which means automatic selection of amino acid substitution model and gamma distributed substitution rate among sites, other parameters were default; (iii) By the program FastME 2.1.5 [5] with default parameters, which were the model LG for computing sequence-to-sequence distances and the BioNJ algorithm for tree inference.
3. The taxonomic tree of the species of the set was obtained from the NCBI Taxonomy database. This tree is not completely resolved (*i.e.*, not binary, some nodes are multifurcating), see Table S3.
4. With the program ASTRAL [6], the version that can satisfy user constraints (available from https://github.com/maryamrabiee/Constrained-search), the consensus of all inferred trees was built, with the taxonomic tree as the constraint. This consensus was used as the reference tree for the species set.

The reference trees are available from https://mouse.belozersky.msu.ru/tools/phylobench/data/ .

Table S3. Number of inner branches in taxonomic trees for species sets. For completely resolved trees of 60 taxa there would be 57 inner branches

ID	# of inner branches	ID	# of inner branches	ID	# of inner branches
AC	16	CH	42	MA	43
AG	29	EB	40	OB	20
AR	33	FB	31	PB	30
AS	36	FI	20	ST	37

Algorithm for the combined sets

The combined sets of alignments are based on 649 archaeal, 650 bacterial, and 650 eukaryotic OGs of 60 protein domains each. We took all archaeal alignments from two sets, AG and AR. From Bacteria and Eukaryota, alignments were selected randomly in two steps. First, we selected 650 random Pfam families from all families represented in bacterial (eukaryotic) OGs. Second, for each selected Pfam family we took a random OG from all OGs from this family in all bacterial (eukaryotic) taxonomic sets. From each of the obtained 1949 OGs we took the correspondent 15-, 30- and 45-sequence alignments. Thus we obtained three combined sets of 1949 alignments each.

Implementation of the algorithms

All the above-mentioned algorithms were implemented in Python 2.7 or Bash scripts available at https://github.com/belozersky321/PhyloBench .

Versions and options of phylogenetic programs

In this study we used the following phylogenetic programs.

- TNT [3] is an implementation of the maximum parsimony algorithm. TNT is being made available with the sponsorship of the Willi Hennig Society. The TNT command *mult* was used with default parameters, which are 10 random addition sequences followed by a tree bisection-reconnection search.
- RAxML 8.2.12 [4] is an implementation of the maximum likelihood algorithm. The following variants of command line parameters were used:
  -m PROTCATJTT
  -m PROTCATAUTO
  -m PROTCATJTT -V
  -m PROTCATWAG -V
  -m PROTCATAUTO -V
  -m PROTGAMMAJTT
  -m PROTGAMMAAUTO
  The parameter -m and the option -V determine the model used for likelihood

calculations. Here “PROT” means that input is a protein alignment, “CAT” (if the option -V is not specified) means dividing alignment columns into a number of categories (by default 25 categories) of evolutionary rates, “GAMMA” means using gamma-distribution of evolutionary rates among columns, the option “-V” (with “CAT” in the value of “-m”) means equal evolutionary rates, “JTT” means using Jones–Taylor–Thornton [7] model of amino acid substitutions, “WAG” means using WAG [8] model of amino acid substitutions, and “AUTO” means the choice of a model from a list (which includes JTT and WAG) according to the likelihood of a draft tree. We always used 12345 as the seed for the random number generator. All other parameters were taken by default.

- IQ-TREE 1.6.12 [9] is an implementation of the maximum likelihood algorithm. We used it with defaults, the command line was:

  ```
  iqtree -s alignment.phy
  ```

  where `alignment.phy` is the name of the file with input alignment. The default behaviour assumes a choice of an evolutionary model among 546 variants according to BIC criterion.

- MrBayes [10] is an implementation of the so-called bayesian method, which is Monte-Carlo search of the tree with maximum *a posterior* probability (MAP tree). This program has plenty of parameters. We varied only the number of generations (the trajectory length) and the frequency of samples. For the trajectory length we used the values 1000, 2000, 4000, 8000, 16000 and 32000, and the frequency of samples was always set to 1/1000 of the trajectory length. Thus, we used the following commands:

  ```
  mcmc ngen=1000 samplefreq=1
  mcmc ngen=2000 samplefreq=2
  mcmc ngen=4000 samplefreq=4
  ```

  and so on. For each trajectory length we tested two trees, one is the MAP tree, and the second is the ASTRAL [6] consensus of the samples (the consensus tree). We used the ASTRAL consensus and not the greedy consensus provided by the program MrBayes itself, because the latter can be not completely resolved. In this work we restrict ourselves to completely resolved (binary) trees only. In MrBayes we always used the

Jones model of amino acid substitutions, i.e., the command

```
prset aamodelpr=fixed(jones)
```

All other parameters were default, for details see MrBayes manual: https://github.com/NBISweden/MrBayes/blob/develop/doc/manual/Manual_MrBayes_v3.2.pdf . In particular, the burnit (the percent of trajectory that was not sampled) was 25% and no rate heterogeneity among sites was used.

- PQ [11] implements an original algorithm. It was used with matrix BLOSUM62 and subtree pruning and regrafting (SPR) search, other parameters by default.
- TREE-PUZZLE [12] implements the quartet-puzzle algorithm. This program often produces not completely resolved trees thus we used it in the following manner. TREE-PUZZLE was run with 100 puzzling steps and with the option “List puzzling step trees → Unique topologies”. The output file “outptorder” contains topologies of 100 trees obtained by shuffling the input order of the sequences. This file (after a proper formatting) is used as input for ASTRAL to obtain a consensus tree.
- FastME 2.1.5 [5] implements several distance measures and a number of distance phylogenetic algorithms. We used the following variants:
  - Variants of sequence-to-sequence distances: p-distance and the distances calculated with amino acid substitution models LG [13] (default) and JTT [7], with or without (default) gamma distribution of rates among sites. Command-line options were:

    – none for LG model without gamma distribution

    – “-pP” for p-distance

    – “-pJ” for JTT model without gamma distribution

    – “-g” for LG model with gamma distribution

    – “-pJ -g” for JTT model with gamma distribution
  - Variants of tree building algorithms: BioNJ (default), Neighbor-joining (NJ), Unweighted neighbor-joining (UNJ), ordinary least squares (OLS) with nearest neighbor interchange (NNI) search, balanced minimum evolution (BME) with NNI search, BME with SPR search. For NNI and SPR searches the starting tree was built with the greedy (TaxAdd) algorithm using OLS or BME criterion. Command-line options were:

– none for BioNJ

– “-m N” for NJ

– “-m U” for UNJ

– “-m O -n O” for OLS with NNI search

– “-m B -n” for BME with NNI search

– “-m B -s” for BME with SPR search

Tree-to-tree distance measures

The following seven measures were tested:

- Robinson–Foulds (RF) distance [14]: the fraction of different splits in two trees;
- A modified Robinson–Foulds distance (MRF). It equals 1 minus the sum of Jaccard measures between best corresponding splits divided by the number of non-trivial splits, see details in https://mouse.belozersky.msu.ru/tools/phylobench/mrf-dist.html ;
- L_1-distance or node distance [15]: the sum of absolute values of differences between path lengths, for all pairs of leaves. Path length is the number of nodes on the path from one leaf to another through a certain tree;
- L_2-distance or path difference metric [16], computed analogously to the L_1-distance but using squares of differences instead of absolute values;
- Quartet distance [17]: the fraction of quartets of leaves with different topology in two trees;
- Agreement distance [18] computed from the size of the unrooted maximum agreement subtree (UMAST) of two trees as the number of leaves that are not present in the UMAST;
- Agreement L-distance [19]: an improvement of the agreement distance.

The agreement L-distance was computed as follows: for two unrooted trees, A and B, with equal sets of leaves, let C be their UMAST. For each leaf that is not present in C, append it to C according to two different topologies, A and B, and compute the number of nodes in C between two alternative positions of this leaf. The agreement L-distance is the sum of these numbers for all extra leaves. We implemented both agreement distances according to the algorithm described by Dong and Kraemer [20], the code in ANSI C is available at

https://github.com/belozersky321/UMAST . Implementation of other distance measures are available at https://github.com/belozersky321/TreeDist .

RESULTS: ADDITIONAL TABLES

Comparison of IQ-TREE with defaults with RAxML with different models

Table S4. Z-scores for pairwise comparisons of IQ-TREE with default parameters with six variants of RAxML. A negative value indicates the superiority of IQ-TREE. In each cell, the top number refers to the 15-sequence alignment set, the middle number refers to the 30-sequence set, and the bottom number refers to the 45-sequence set.

Substitution matrix for RAxML	JTT			AUTO		
Rate het. for RAxML	None	CAT	Gamma	None	CAT	Gamma
Z-score	2.3 3.3 1.3	0.33 1.7 0.67	0.73 1.5 –0.36	–0.65 0.77 1.06	–2.73 0.27 –1.7	–2.0 –1.3 –2.1

Comparison of MCMC trajectory lengths in MrBayes

Table S5. Z-scores for pairwise comparisons of MAP and consensus trees built by MrBayes with different trajectory lengths, with each other and with RAxML, on three combined sets. A negative value indicates the superiority of the model in the row. In each cell, the top number refers to the 15-sequence alignment set, the middle number refers to the 30-sequence set, and the bottom number refers to the 45-sequence set. For RAxML, the JTT model of amino acid substitutions with no rate heterogeneity was used. See the section "Versions and options of phylogenetic programs" of this document for parameters of MrBayes.

		MAP						Consensus						RAxML
	Steps	1K	2K	4K	8K	16K	32K	1K	2K	4K	8K	16K	32K	
MAP	1K		−0.54 15.9 37.6	0.15 19.9 47.2	−0.28 21.3 49.3	0.02 21.3 48.9	0.25 20.9 48.9	6.68 26.1 33.4	3.64 30.8 56.4	2.95 28.9 57.7	3.33 26.9 55.5	2.79 26.3 54.4	2.88 26.3 53.4	−2.13 20.9 49.0
	2K	0.54 −15.9 −37.6		0.89 5.76 16.9	0.36 7.67 21.5	0.76 7.71 21.7	1.05 7.26 21.3	6.57 6.70 −12.8	6.57 22.1 30.3	4.50 17.3 33.2	4.94 15.2 31.2	4.32 14.5 30.3	4.36 14.2 28.9	−1.89 7.40 21.8
	4K	−0.15 −19.9 −47.2	−0.89 −5.76 −16.9		−0.61 1.80 6.30	−0.18 1.72 6.90	0.15 1.29 6.37	5.86 1.42 −26.2	4.55 13.7 9.86	4.49 14.1 24.3	4.34 10.4 20.7	3.70 9.80 19.2	3.69 9.68 17.5	−2.60 1.98 7.57
	8K	0.28 −21.3 −49.3	−0.36 −7.67 −21.5	0.61 −1.80 −6.30		0.52 −0.14 0.44	0.89 −0.60 −0.29	6.40 0.05 −29.3	5.23 12.6 4.25	4.48 11.9 16.6	5.62 10.8 18.5	4.50 9.42 14.3	4.64 9.13 12.9	−2.20 0.35 1.64
	16K	−0.02 −21.3 −48.9	−0.76 −7.71 −21.7	0.18 −1.72 −6.90	−0.52 0.14 −0.44		0.44 −0.49 −0.84	5.96 0.14 −29.1	4.78 12.7 4.00	4.15 12.2 17.0	4.96 10.3 17.2	4.58 10.3 16.3	4.32 9.65 14.1	−2.53 0.47 1.31
	32K	−0.25 −20.9 −48.9	−1.05 −7.26 −21.3	0.15 −1.29 −6.37	−0.89 0.60 0.29	−0.44 0.49 0.84		5.93 0.43 −28.8	4.51 13.4 4.61	3.93 12.7 18.1	4.91 11.5 18.6	4.15 11.1 16.9	4.44 11.8 16.5	−2.76 0.89 2.09
Consensus	1K	−6.68 −26.1 −33.4	−6.57 −6.70 12.8	−5.86 −1.42 26.2	−6.40 −0.05 29.3	−5.96 −0.14 29.1	−5.93 −0.43 28.8		−2.33 11.0 37.3	−3.15 8.68 39.5	−2.79 6.47 37.1	−3.44 5.98 36.1	−3.38 5.82 35.0	−7.26 0.21 29.5
	2K	−3.64 −30.8 −56.4	−6.57 −22.1 −30.3	−4.55 −13.7 −9.86	−5.23 −12.6 −4.25	−4.78 −12.7 −4.00	−4.51 −13.4 −4.61	2.33 −11.0 −37.3		−1.14 −2.86 10.6	−0.63 −5.46 8.78	−1.45 −5.99 6.99	−1.41 6.09 5.68	−6.09 −11.5 −2.82
	4K	−2.95 −28.9 −57.7	−4.50 −17.3 −33.2	−4.49 −14.1 −24.3	−4.48 −11.9 −16.6	−4.15 −12.2 −17.0	−3.93 −12.7 −18.1	3.15 −8.68 −39.5	1.14 2.86 −10.6		0.59 −3.40 −1.73	−0.38 −4.22 −4.33	−0.28 −4.41 −5.77	−5.44 −10.1 −15.0
	8K	−3.33 −26.9 −55.5	−4.94 −15.2 −31.2	−4.34 −10.4 −20.7	−5.62 −10.8 −18.5	−4.96 −10.3 −17.2	−4.91 −11.5 −18.6	2.79 −6.47 −37.1	0.63 5.46 −8.78	−0.59 3.40 1.73		−1.17 −1.11 −3.22	−1.10 −1.40 −5.26	−5.83 −8.12 −14.9
	16K	−2.79 −26.3 −54.4	−4.32 −14.5 −30.3	−3.70 −9.80 −19.2	−4.50 −9.42 −14.3	−4.58 −10.3 −16.3	−4.15 −11.1 −16.9	3.44 −5.98 −36.1	1.45 5.99 −6.99	0.38 4.22 4.33	1.17 1.11 3.22		0.12 −0.32 −2.41	−5.35 −7.59 −13.0
	32K	−2.88 −26.3 −53.4	−4.36 −14.2 −28.9	−3.69 −9.68 −17.5	−4.64 −9.13 −12.9	−4.32 −9.65 −14.1	−4.44 −11.8 −16.5	3.38 −5.82 −35.0	1.41 6.09 −5.68	0.28 4.41 5.77	1.10 1.40 5.26	−0.12 0.32 2.41		−5.39 −7.65 −11.8
RAxML		2.13 −20.9 −49.0	1.89 −7.40 −21.8	2.69 −1.98 −7.57	2.20 −0.35 −1.64	2.53 −0.47 −1.31	2.76 −0.89 −2.09	7.26 −0.21 −29.5	6.09 11.5 2.82	5.44 10.1 15.0	5.83 8.12 14.9	5.35 7.59 13.0	5.39 7.65 11.8	

Comparison of different distance algorithms implemented in FastME

Table S5. Z-scores for pairwise comparisons of six distance algorithms and mean RF distances from inferred to reference trees. NJ refers to neighbor-joining, UNJ refers to unweighted neighbor-joining, BME (NNI) and BME (SPR) refer, respectively, to NNI and SPR search with the balanced minimum evolution criterion, OLS refers to NNI search with the ordinary least squares criterion. In each cell, the top number refers to the 15-sequence set, the middle number refers to the 30-sequence set, and the bottom number refers to the 45-sequence set. Parameters for measuring sequence-to-sequence distances were default, i.e., the LG model with equal substitution rates among sites.

	BioNJ	NJ	UNJ	OLS	BME (NNI)	BME (SPR)	Mean distance
BioNJ		–1.61 –0.03 –1.99	–3.00 –3.35 –6.40	–2.68 –5.22 –4.28	0.17 1.18 4.32	0.78 2.58 7.80	0.5217 0.5771 0.6071
NJ	1.61 0.03 1.99		–1.70 –3.29 –4.75	–1.60 –5.11 –3.10	1.46 1.21 5.60	2.17 2.64 9.01	0.5235 0.5771 0.6091
UNJ	3.00 3.35 6.40	1.70 3.29 4.75		–0.41 –2.94 0.22	2.48 3.49 8.26	3.05 4.84 11.3	0.5255 0.5800 0.6115
OLS	2.68 5.22 4.28	1.69 5.11 3.10	0.41 2.94 –0.22		2.73 5.88 8.14	3.21 7.16 10.4	0.5261 0.5831 0.6127
BME (NNI)	–0.17 –1.18 –4.32	–1.46 –1.21 –5.60	–2.48 –3.49 –8.26	–2.73 –5.88 –8.14		1.24 2.44 3.67	0.5215 0.5757 0.6035
BME (SPR)	–0.78 –2.58 –7.80	–2.17 –2.64 –9.01	–3.05 –4.84 –11.3	–3.21 –7.16 –10.4	–1.24 –2.44 –3.67		0.5206 0.5741 0.6023

Comparison of six methods using NCBI taxonomy as reference

Table S6. Z-scores for the comparison of six methods on the three combined sets and the mean RF distances from the references, using NCBI taxonomy as reference. Denotations are the same as in Table 4 of the main text, in particular, the top number refers to the 15-sequence set, the middle number refers to the 30-sequence set, and the bottom number refers to the 45-sequence set. All methods were run with optimal parameters (see description to Table 4 in the main text).

	TNT	RAxML	MrBayes	TREE-PUZZLE	PQ	FastME	Mean distance
TNT		10.9 15.5 18.2	15.7 23.5 27.5	16.6 25.8 28.6	17.7 24.8 27.6	18.4 28.4 33.8	0.6017 0.6676 0.7088
RAxML	−10.9 −15.5 −18.2		5.87 10.0 11.3	6.80 13.0 11.8	8.03 11.4 11.3	7.72 14.6 16.7	0.5739 0.6408 0.6813
MrBayes	−15.7 −23.5 −27.5	−5.87 −10.0 −11.3		2.41 5.16 3.17	4.15 4.37 3.13	3.64 6.68 8.10	0.5635 0.6282 0.6701
TREE-PUZZLE	−16.6 −25.8 −28.6	−6.80 −13.0 −11.8	−2.41 −5.16 −3.17		2.27 −0.74 0.14	1.21 1.28 4.97	0.5587 0.6211 0.6664
PQ	−17.7 −24.8 −27.6	−8.03 −11.4 −11.3	−4.15 −4.37 −3.13	−2.27 0.74 −0.14		−1.07 1.89 4.44	0.5544 0.6220 0.6662
FastME	−18.4 −28.4 −33.8	−7.72 −14.6 −16.7	−3.64 −6.68 −8.10	−1.21 −1.28 −4.97	1.07 −1.89 −4.44		0.5566 0.6195 0.6614

Comparison of six methods using consensus without NCBI constrains as reference

Table S7. Z-scores for the comparison of six methods on the three combined sets and the mean RF distances from the references, using ASTRAL consensus of trees inferred from all OGs with three methods, as reference. Denotations are the same as in Table 4 of the main text and in Table S6.

	TNT	RAxML	MrBayes	TREE-PUZZLE	PQ	FastME	Mean distance
TNT		10.2 16.1 19.6	14.5 26.3 33.2	17.8 29.8 34.4	19.7 30.9 37.3	19.5 31.8 41.2	0.5568 0.6187 0.6491
RAxML	−10.2 −16.1 −19.6		6.72 12.0 16.1	9.29 15.7 16.2	11.1 17.2 20.2	10.1 17.5 23.1	0.5287 0.5869 0.6164
MrBayes	–14.5 −26.3 −33.2	–6.72 −12.0 −16.1		4.86 6.28 4.43	7.23 9.21 9.68	5.81 7.80 11.6	0.5167 0.5697 0.5983
TREE-PUZZLE	−17.8 −29.8 −34.4	−9.29 −15.7 −16.2	−4.86 −6.28 −4.43		3.56 4.27 6.88	0.88 1.17 7.09	0.5056 0.5596 0.5923
PQ	−19.7 −30.9 −37.3	−11.1 −17.2 −20.2	−7.23 −9.21 −9.68	−3.56 −4.27 −6.88		–2.56 –2.87 0.058	0.4982 0.5535 0.5839
FastME	−19.5 −31.8 −41.2	−10.1 −17.5 −23.1	−5.81 −7.80 −11.6	−0.88 −1.17 −7.09	2.56 2.87 −0.058		0.5039 0.5580 0.5839

Separate results for three parts of the combined sets

Table S8. Z-scores for the comparison of six methods and the mean RF distances from the references, on three sets of archaeal alignments (AR + AG together, 15- 30- and 45-sequence alignments). Denotations are the same as in Table 4 of the main text and in Tables S6 and S7.

	TNT	RAxML	MrBayes	TREE-PUZZLE	PQ	FastME	Mean distance
TNT		6.25 8.84 8.72	9.88 14.0 16.6	10.9 18.2 22.6	12.7 19.7 23.3	12.4 18.6 23.9	0.5012 0.5554 0.5922
RAxML	−6.25 −8.84 −8.72		4.53 6.65 10.4	5.47 11.2 16.0	7.15 12.1 16.0	6.42 10.6 18.0	0.4703 0.5267 0.5692
MrBayes	−9.88 −14.0 −16.6	−4.53 −6.65 −10.4		2.24 6.47 8.95	4.74 8.17 9.53	3.6 5.72 10.2	0.4556 0.5112 0.5497
TREE-PUZZLE	−10.6 −18.2 −22.6	−5.47 −11.2 −16.0	−2.24 −6.47 −8.95		3.02 2.58 1.54	1.12 −1.50 −0.06	0.4471 0.4936 0.5293
PQ	−12.7 −19.7 −23.3	−7.15 −12.1 −16.0	−4.74 −8.17 −9.53	−3.02 −2.58 −1.54		−1.94 −3.82 −1.54	0.4364 0.4874 0.5262
FastME	−12.4 −18.6 −23.9	−6.42 −10.6 −16.0	−3.6 −5.72 −9.53	−1.12 1.50 −1.54	1.94 3.82 1.54		0.4435 0.4972 0.5295

Table S9. Z-scores for the comparison of six methods and the mean RF distances from the references, on three sets of bacterial alignments (bacterial part of 15- 30- and 45-sequence combined sets). Denotations are the same as in Table 4 of the main text and in Tables S6– S8.

	TNT	RAxML	MrBayes	TREE-PUZZLE	PQ	FastME	Mean distance
TNT		5.82 8.10 9.61	9.03 13.1 17.9	9.96 14.6 18.4	11.1 15.5 20.5	10.2 15.9 21.1	0.6318 0.6923 0.7186
RAxML	−5.82 −8.10 −9.61		4.22 6.53 9.84	4.72 7.74 10.5	5.89 9.23 14.1	4.65 8.71 13.4	0.6056 0.6674 0.6948
MrBayes	−9.03 −13.1 −17.9	−4.22 −6.53 −9.84		1.56 3.22 3.71	3.13 5.34 8.11	1.40 4.04 7.22	0.5915 0.6529 0.6781
TREE-PUZZLE	−9.96 −14.6 −18.4	−4.72 −7.74 −10.5	−1.56 −3.22 −3.71		1.90 2.65 5.36	–0.29 0.72 2.91	0.5855 0.6450 0.6702
PQ	−11.1 −15.5 −20.5	−5.89 −9.23 −14.1	−3.13 −5.34 −8.11	−1.90 –2.65 −5.36		−2.18 –1.76 –2.43	0.5788 0.6391 0.6601
FastME	−10.2 −15.9 −21.1	−4.65 −8.71 −13.4	−1.40 −4.04 −7.22	0.29 −0.72 −2.71	2.18 1.76 2.43		0.5864 0.6434 0.6648

Table S10. Z-scores for the comparison of six methods and the mean RF distances from the references, on three sets of eukaryotic alignments (eukaryotic part of 15- 30- and 45-sequence combined sets). Denotations are the same as in Table 4 of the main text and in Tables S6– S9.

	TNT	RAxML	MrBayes	TREE-PUZZLE	PQ	FastME	Mean distance
TNT		7.51 11.2 15.2	10.4 16.9 21.2	10.0 15.9 17.9	9.43 14.4 17.7	11.5 18.8 25.0	0.5792 0.6443 0.6796
RAxML	−7.51 −11.2 −15.2		3.83 6.71 6.19	3.25 5.21 2.05	2.40 4.18 2.46	3.97 2.92 8.82	0.5431 0.6035 0.6317
MrBayes	−10.4 −16.9 −21.2	−3.83 −6.71 −6.19		0.30 –0.52 –3.32	–0.42 –0.85 –6.95	1.13 2.03 4.11	0.5303 0.5863 0.6189
TREE-PUZZLE	−10.0 −15.9 −17.9	−3.25 −5.21 −2.05	−0.30 0.52 3.32		–0.84 −0.45 0.7	0.95 2.87 8.46	0.5291 0.5878 0.6265
PQ	−9.43 −14.4 −17.7	−2.40 −4.18 −2.46	0.42 0.85 2.44	0.84 0.45 −0.7		1.61 2.92 6.95	0.5321 0.5889 0.6251
FastME	−11.5 −18.8 −25.0	−3.97 −7.89 −8.82	−1.13 −2.03 −4.11	−0.95 −2.87 −8.46	–1.61 −2.92 −6.95		0.5258 0.5809 0.6095

REFERENCES

1. Mistry J., Chuguransky S., Williams L., Qureshi M., Salazar G.A., Sonnhammer E.L.L., Tosatto S.C.E., Paladin L., Raj S., Richardson L.J., Finn R.D., Bateman A. 2021. Pfam: The protein families database in 2021. Nucleic Acids Res. 49(D1):D412–D419.
2. Edgar R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5):1792–1797.
3. Goloboff P., Catalano S. 2016. TNT, version 1.5, including a full implementation of phylogenetic morphometrics. Cladistics. DOI 10.1111/cla.12160.
4. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30:1312–1313.
5. Lefort V., Desper R., Gascuel O. 2015. FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program. Mol. Biol. Evol., 32:2798–2800, 2015.
6. Zhang C., Rabiee M., Sayyari E., Mirarab S. ASTRALIII: Polynomial Time Species Tree Reconstruction from Partially Resolved Gene Trees. BMC Bioinformatics, 19(S6):153, 2018.
7. Jones D., Taylor W., Thornton J. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci., 8:275–282, 1992.
8. Whelan S, Goldman N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol., 18:691–699, 2001.
9. Nguyen L.-T., Schmidt H.A., von Haeseler A., Minh B.Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies.. Mol. Biol. Evol., 32:268-274, 2015.
10. Ronquist F., Teslenko M., van der Mark P. et al. MRBAYES 3.2: Efficient Bayesian phylogenetic inference and model selection across a large model space. Syst. Biol., 61:539–542, 2012.
11. Penzar D., Krivozubov M., Spirin S. PQ, a new program for phylogeny reconstruction. BMC Bioinformatics, 19:374, 2018.
12. Schmidt H.A., Strimmer K., Vingron M., von Haeseler A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics, 18:502–504, 2002.
13. LG

14. Robinson D.R., Foulds L.R. Comparison of phylogenetic trees. Mathematical Biosciences, 53(1–2):131–147, 1981.
15. Williams W.T., Clifford H.T. On the comparison of two classifications of the same set of elements. Taxon, 20:519–522, 1971.
16. Penny D., Foulds L.R., Hendy M.D. Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences. Nature, 297:197–200, 1982.
17. Estabrook G.F., McMorris F.R., Meacham C.A. 1985. Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. Systematic Biology, 34(2):193–200, 1985.
18. Gordon A.D. On the assessment and comparison of classifications. In Tomassine R. editor. Analyse de données et informatique, Le Chesnay, France: INRIA, pages 149–160, 1980.
19. Goddard W., Kubicka E., Kubicki G., McMorris F.R. The agreement metric for labeled binary trees. Mathematical Biosciences, 123:215–226, 1994.
20. Dong S., Kraemer E. 2004. Calculation, Visualization, and Manipulation of MASTs (Maximum Agreement Subtrees). In: Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB 2004).