PhyloBench: A Benchmark for Evaluating Phylogenetic Programs

Sergey Spirin [1,2,*] Andrey Sigorskikh,[3] Aleksei Efremov,[3] Dmitry Penzar,[3,4] and Anna Karyagina[1,5,6]

[1]Belozersky Institute, Lomonosov Moscow State University, Moscow, Russia
[2]Higher School of Economics, Moscow, Russia
[3]Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia
[4]Artificial Intelligence Research Institute, Moscow, Russia
[5]Gamaleya Center of Epidemiology and Microbiology, Moscow, Russia
[6]Institute of Agricultural Biotechnology, Moscow, Russia

***Corresponding author**: E-mail: sas@belozersky.msu.ru.

Associate editor: Andrey Rzhetsky

Abstract

Phylogenetic inference based on protein sequence alignment is a widely used procedure. Numerous phylogenetic algorithms have been developed, most of which have many parameters and options. Choosing a program, options, and parameters can be a nontrivial task. No benchmark for comparison of phylogenetic programs on real protein sequences was publicly available. We have developed PhyloBench, a benchmark for evaluating the quality of phylogenetic inference, and used it to test a number of popular phylogenetic programs. PhyloBench is based on natural, not simulated, protein sequences of orthologous evolutionary domains. The measure of accuracy of an inferred tree is its distance to the corresponding species tree. A number of tree-to-tree distance measures were tested. The most reliable results were obtained using the Robinson–Foulds distance. Our results confirmed recent findings that distance methods are more accurate than maximum likelihood (ML) and maximum parsimony. We tested the bayesian program MrBayes on natural protein sequences and found that, on our datasets, it performs better than ML, but worse than distance methods. Of the methods we tested, the Balanced Minimum Evolution method implemented in FastME yielded the best results on our material. Alignments and reference species trees are available at https://mouse.belozersky.msu.ru/tools/phylobench/ together with a web-interface that allows for a semi-automatic comparison of a user's method with a number of popular programs.

Key words: **phylogenetic inference, benchmark.**

Introduction

There are a number of algorithms and programs available for inferring protein phylogeny from a multiple sequence alignment. Most programs allow users to select several parameters, with variations in programs and parameters often affecting the results. An evaluation of the relative accuracy of programs, or the choice of options, is often based on simulated sequence alignments (Wu et al. 2012; Hollich et al. 2005). In several studies where empirical alignments were used, such as (Zhou et al. 2017), the evaluation was performed according to likelihoods of inferred trees.

There are few studies where comparisons of phylogenetic programs were performed on alignments of real sequences of orthologous proteins and the accuracy was evaluated according to proximity to a species tree (Krivozubov and Spirin 2010; Gonnet 2012; Penzar et al. 2018). These studies provided significantly different results from those made on simulations. In particular, in these studies phylogenetic reconstructions performed with distance methods often turned to be more accurate than results of maximum likelihood. This suggests that comparisons made with simulated alignments may not be relevant for real data. No benchmark for evaluation of phylogenetic programs on real sequences has, thus far, been made publicly available.

Therefore, the aim of the present work was to create a benchmark and workflow that allow evaluation of phylogenetic programs on large sets of natural protein sequence alignments. Our approach comprised selecting a number of orthologous groups (OGs) of proteins and to compare trees inferred from their alignments with corresponding species trees. To avoid problems with domain shuffling, we used evolutionary domains according to Pfam (Mistry et al. 2021) rather than full-length protein sequences. We prepared 12 sets of protein alignments, 60 sequences in each alignment, each set represents proteins from one large taxon (Metazoa, Actinobacteria, etc.).

Received: November 17, 2023. **Revised:** April 05, 2024. **Accepted:** April 22, 2024

Resources

Evaluation of phylogenetic programs were performed on random selections of 15, 30, and 45 proteins from each alignment, to avoid biases related to the same topology of reference (*i.e.* species) trees for different alignments.

Surely, a species tree not always exactly reflects the real phylogeny of an OG of proteins. Differences may arise due to horizontal gene transfer, errors in selecting orthologs, or incomplete linear sorting. However, we may expect that the differences are not large enough to prevent a fair comparison of phylogenetic methods. On average, discordances between real protein trees and the corresponding species trees would increase distances from different phylogenetic reconstructions to the species trees in the same degree. Thus, the distances from the species tree to better reconstructions should remain, on average, less than the distances to worse reconstructions. Such discordances cannot cause a wrong signal, they can just introduce some noise making the observed signal less visible.

To test whether the mentioned noise would prevent the comparison of the methods, we performed computational experiments. Namely, we compared phylogenetic reconstructions performed with one method but on two series of sequence alignments, first, intact empirical alignments, and second, the same alignments with one-fifth of information removed. This test, for all alignment sets, showed statistically significant advantages of reconstructions from intact alignments. Thus we demonstrated an applicability of our benchmark to comparison of phylogenetic methods. The same experiment was used to identify the tree-to-tree distance measure providing the best statistical significance in comparison of the reconstructions.

Using the benchmark, we compared a number of popular phylogenetic tools and their parameters. An advantage of distance methods was confirmed. Also, the tests of a bayesian method showed that a consensus of the output ensemble of trees is on average closer to the reference than the result of the ML method, while farther than the result of a distance method.

In the current work, we focused mainly on investigating the quality of phylogenetic inference based on a restricted phylogenetic signal, *i.e.* on a set of rather short homologous protein sequences. The issue of discordant signals due to horizontal gene transfer, recombination, loss of paralogs, incomplete lineage sorting, *etc.* remains beyond the scope of this study. However, the situation where phylogenetic reconstruction has to be done on the basis of one relatively short alignment is fairly typical. For example, that constantly occurs in studies of the evolution of paralogous families. We hope that our benchmark would help in adequate choice of programs and parameters for such studies. Also, it can be helpful for software developers.

Results and Discussion

Benchmark for Testing Phylogenetic Methods

Twelve sets of 60 living species each were selected. Two sets, AG and AR, represent Archaea, four sets, AC, FI, OB, and PB, are from different taxa of Bacteria, other six sets, AS, CH, EB, FB, MA, and ST, represent Eukaryota. See supplementary tables S1 and S2, Supplementary Material online for characteristics of the species sets and the supplementary `Species.xlsx`, Supplementary Material online for the lists of species in each set.

From each species set, the maximum possible number of OGs of Pfam evolutionary domains were selected. In total, 17,446 OGs were formed, with 649 representing Archaea, 2,522 Bacteria, and 14,275 Eukaryota. Sequences of these domains were aligned and used for constructing reference trees (see Materials and Methods). From each OG, three sequence subsets were randomly and independently selected; the first subset consisting of 15, the second of 30 and the third of 45 sequences. This was done to make tree topologies more diverse. Alignments of these sequence subsets form 36 (three sizes times 12 species sets) taxonomic sets of alignments. Comparison of trees inferred from these alignments with the reference trees can be used for estimation of the quality of an inference procedure.

To make comparisons faster and to avoid a skew towards eukaryotes, we extracted three sets of alignments from the taxonomic sets where three domains of cellular organisms were represented equally. We called them combined sets. The main part of PhyloBench is three combined sets of 15-sequence, 30-sequence, and 45-sequence alignments, together with reference trees for them. Each combined set consists of 649 archaeal, 650 bacterial, and 650 eukaryotic alignments. These alignments and trees are available for download (see subsection "Online service"). The distributions of lengths of alignments from the 45-sequence combined set are shown in Fig. 1.

The Robinson–Foulds Distance was Best for Comparison of Phylogenetic Methods

We compared 10 variants of measuring the relative accuracy of tree inference. There were two variants of reference trees, namely the taxonomy trees, which are the unresolved trees derived from NCBI taxonomy, and resolved species trees, which were constructed based on trees inferred with three programs on all available OGs for a given species set (see Materials and Methods). We tested seven tree-to-tree distance measures for the resolved trees and three measures, Robinson–Foulds (RF) distance (Robinson and Foulds 1981), Quartet (Q) distance (Williams and Clifford 1971) and an original modified Robinson–Foulds distance (MRF) for measuring distances between resolved trees inferred by TNT from intact and damaged alignments (see Materials and Methods) and unresolved taxonomic trees. We did not test L_1, L_2, agreement (A) and agreement L- (AL) distances to unresolved taxonomic trees because it was not clear how to correctly generalize these measures to unresolved trees. For trees inferred from intact and damaged alignments, we calculated Z-score for differences between distances from both trees to the reference tree. The idea was that more sensitive is the distance measure to errors in phylogenetic reconstruction, the higher this Z-score should be.

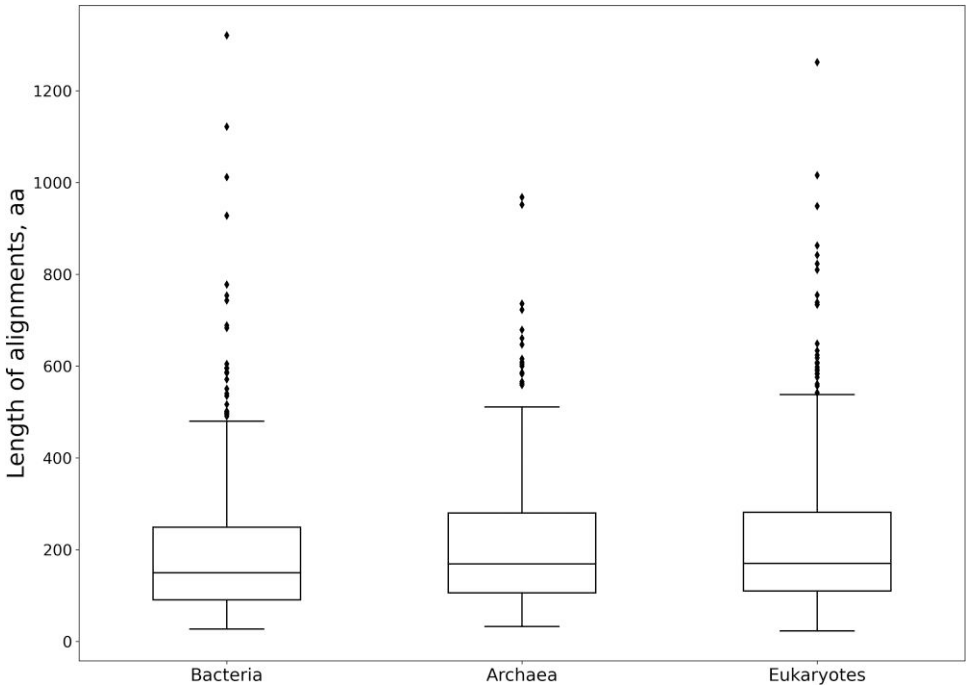

Fig. 1. Box plots for the length distributions of the 45-sequence alignments from the combined set. Lines in boxes are for lower quartiles, medians, and upper quartiles. The whiskers mark maximum and minimum within the 1.5 interquartile range from the quartiles. Outliers are plotted with black diamonds.

The results of testing the distance measures on 12 taxonomic sets of 45-sequence alignments are given in the supplementary table S4, Supplementary Material online. In seven out of 12 cases, the largest median (over five variants of alignment damaging) Z-score was obtained using the RF distance to the resolved species tree. For three sets, the MRF distance to the resolved species trees provided the best results, and for two sets, the best was the RF distance to the taxonomic tree. The results for 15-sequence and 30-sequence alignments were close, namely in most cases the best was the RF distance to the resolved species trees.

This result suggests that the RF distance to the resolved species trees is typically the most sensitive to removing a part of information from the input data. Based on these results, we decided to use mainly this measure for comparison of phylogenetic methods.

Applicability of Our Species Trees to Benchmarking

For all twelve 45-sequence taxonomic sets, RF distances from the inferred trees to both variants of reference trees reliably distinguished between intact and damaged alignments (Z-test $P < 10^{-5}$). Thus, PhyloBench is suitable for comparing phylogenetic algorithms. This suitability is also valid for the combined sets, as the reference trees are the same for the alignments in these sets as in the taxonomic sets. The following subsections contain results of comparisons performed using RF distances to resolved species trees on three combined sets.

Comparison of Models for Maximum Likelihood

One of the parameters of maximum likelihood (ML) is the probability model for amino acid substitutions. As an implementation of ML, we used RAxML (Stamatakis 2014). Among variants of substitution probabilities available in RAxML, we compared two: the Jones–Taylor–Thornton (JTT) matrix (Jones et al. 1992), and the AUTO option (the automatic choice of a substitution matrix). Another tested parameter concerned the heterogeneity of substitution rates among sites of the input alignment. RAxML provides three rate heterogeneity models: first, the widely used Gamma distribution of rates (Gu et al. 1995), second, the CAT model, which consists of dividing all sites into 25 (by default) categories with different substitution rates, and third, equal substitution rates among sites. Thus, six variants of models were tested: two variants of substitution probabilities times three rate heterogeneity models. Supplementary table S5, Supplementary Material online contains Z-scores for the pairwise comparisons of these six variants on three combined sets and the relative computational times that were required.

Neither automatic choice of substitution model nor taking into account rate heterogeneity showed any significant increase of quality; otherwise, on our data both demonstrated even some (mostly insignificant) decrease. For example, we compared two models of RAxML: first, JTT with no rate heterogeneity (-m PROTCATJTT and -V in the command line of RAxML), and second, JTT with Gamma distributed rates (-m PROTGAMMAJTT). For three combined sets, the first variant was better for 247 15-sequence alignments, 483 30-sequence alignment and 662 45-sequence alignments, while the second variant was better for 206, 407, and 602 alignments, respectively. Summarizing, the first variant was better in 1,392 cases and worse in 1,215 cases. Such results would have an extremely low probability if Gamma had any more or less significant advantage. This was unexpected, as alignment

Table 1. The number of 45-sequence alignments from the combined set with relations between tree likelihoods and tree qualities for the JTT and WAG models

	JTT likelihood higher	WAG likelihood higher
JTT better[a]	**169**	**466**
Equal[b]	**148**	**565**
WAG better[c]	**127**	**474**

Substitution rates were equal among sites. [a]The row "JTT better" contains the number of alignments for which the tree inferred with the JTT model was closer to the reference tree in comparison to the tree inferred using the WAG model. [b]The row "Equal" contains the number of alignments for which two trees either coincided or had equal RF distances to the reference tree. [c]The row "WAG better" contains the number of alignments for which the tree inferred with the WAG model was closer to the reference tree.

positions obviously differ in the rate of substitution accumulations. However, the relative results of the CAT model tended to be better when the number of sequences in the input alignments increased.

Also we tested the General Time Reversible (GTR) model (PROTCATGTR and PROTGAMMAGTR in RAxML, data not shown). As expected, the average accuracy with GTR was worse than with other tested models for 15- and 30-sequence alignments, while for 45-sequence alignments the accuracy was approximately the same. The computation time with GTR was 5 to 10 times longer than with JTT.

Is it Informative to Compare Likelihoods for Different Models?

The automatic detection of a substitution model (the AUTO model option in RAxML) is based on a common belief that for a certain alignment a model can be chosen based on the likelihoods that different models provide for optimal or suboptimal trees. We tested the correspondence between likelihoods of optimized trees and their proximity to reference trees for two substitution models that provide similar quality of inference, namely WAG (Whelan and Goldman 2001) and JTT. The results for the 45-sequence combined set are listed in Table 1. The results for the 30- and 15-sequence sets were similar (data not shown).

The likelihoods for Table 1 were calculated for trees optimized with the corresponding model. Usage of the WAG model resulted in higher likelihoods in the vast majority of cases. At the same time the number of cases when the tree optimized with WAG is closer to the reference tree was approximately the same as the number of opposite cases. This result suggests that the relative quality of phylogenetic inferences based on different models is not strongly connected with difference in likelihoods calculated using these models. This may explain the relatively poor results of the AUTO model.

The program IQ-Tree (Nguyen et al. 2015) provides a possibility to choose between models having different numbers of free parameters, in particular between models with gamma distribution of substitution rates and without it. The choice is performed according to Bayesian information criterion (BIC) taking into account both the

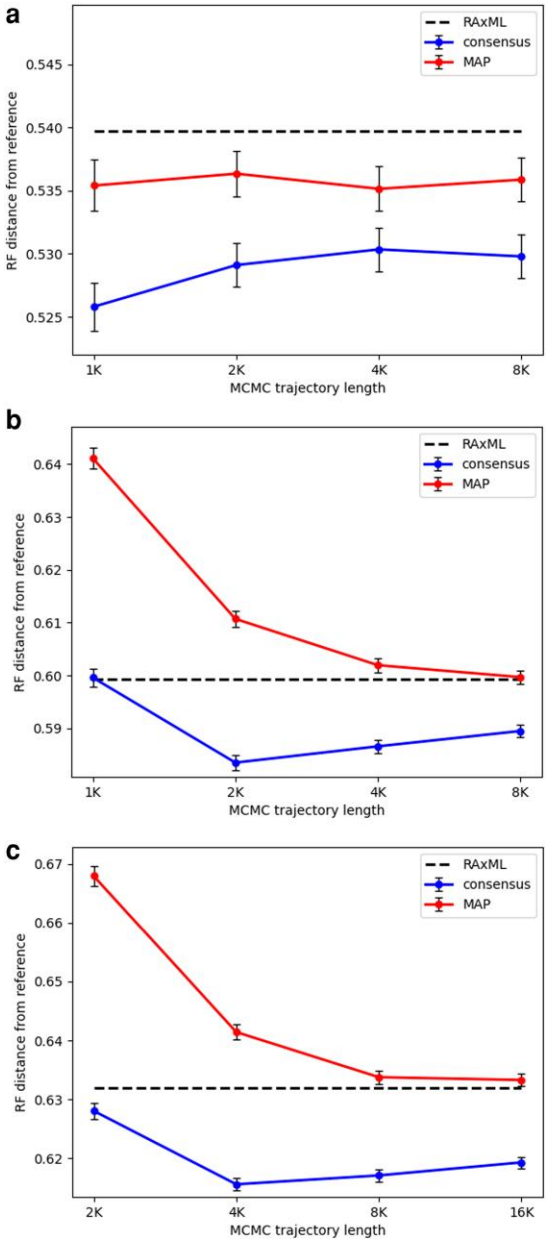

Fig. 2. Average quality of phylogenetic inference by MrBayes in comparison to RAxML, in dependence on MCMC trajectory length. The vertical axis corresponds to average Robinson–Foulds distances from MrBayes and RAxML trees to the resolved species trees. The error bars reflect standard errors of the average differences between two distances from the reference, the distance to a RAxML tree and the distance to a MrBayes tree. Calculations were performed on the combined sets of alignments of 15 (a), 30 (b), and 45 (c) sequences.

number of parameters and likelihoods of a draft tree computed with different models (refer to the paper on ModelFinder (Kalyaanamoorthy et al. 2017) for details). By default, for each alignment IQ-Tree selects between 546 models, which differ in substitution matrices (20 variants), state frequencies, and ways of modeling heterogeneity across sites.

We tested the default behavior of IQ-Tree on three combined sets and compare the results with the results

of different variants of RAxML. The results are available in supplementary table S6, Supplementary Material online. For the 15-sequence and 30-sequence sets, IQ-Tree with defaults was significantly worse than RAxML with the fixed (JTT) substitution matrix and no rate heterogeneity. For the 45-sequence set the superiority of RAxML was not significant ($Z = 1.3$). This result suggests that even simultaneous choice of substitution model and rate heterogeneity model according to BIC does not guarantee an improvement in accuracy.

The reasons for these experimental facts are not clear. We hypothesize that the main source of even worse results of an automatic choice of a model was in the Li–Gascuel model (LG) (Le and Gascuel 2008). This model very often provides the best likelihood, while on our datasets it showed slightly inferior results compared to JTT or WAG. For example, see supplementary table S9, Supplementary Material online with results of FastME; when used in RAxML, the relative results of JTT and LG were similar.

Phylogenetic Inference with MCMC: Evidence in Favor of the Existence of an Optimal Trajectory Length

The MrBayes (Ronquist et al. 2012) program uses the Metropolis–Hastings algorithm, a type of Monte-Carlo Markov Chain (MCMC) search in the tree space with the *a posterior* probability of trees as the objective function. The program outputs both the tree with the maximum *a posterior* probability (the MAP tree) and the ensemble of trees from the MCMC trajectory with their *a posterior* probabilities. We used the ASTRAL (Zhang et al. 2018) consensus of the ensemble as an alternative result of MrBayes.

One of the main parameters of the program is the trajectory length, *i.e.* the number of iterations of the MCMC search. We investigated the dependence of the average quality of two trees, the MAP tree and the consensus tree, on the number of iterations. Figure 2 shows the results of these investigations in comparison with the results of RAxML. Here we used the Jones model in MrBayes and the JTT model in RAxML ("Jones" and "JTT" are actually two names of the same model), with no rate heterogeneity.

MAP trees were, on average, more distant from references than consensus trees for any number of iterations, and the difference was reliable in all cases. For 30-sequence and 45-sequence alignments, the average RF distance from the references to the MAP trees decreased with an increasing number of iterations. For consensus trees, there were optimal numbers of iterations. For example, for 45-sequence alignments, the consensus tree at 4,000 iterations was closer to the reference than the consensus tree at 32,000 iterations in 741 cases out of 1,949 and was more distant in 554 cases, $Z = -5.51$ (the full table of Z-scores see in supplementary table S7, Supplementary Material online). This optimal trajectory length increased with alignment size. The effect of decreasing quality when trajectory length became greater than its optimum may be a result of some kind of overfitting.

For alignments of 30 or 45 sequences, the average distance from the reference trees to the MrBayes MAP trees for any number of iterations was greater than the distance to the RAxML trees. However, for 15 sequences, the MrBayes MAP trees were, on average, closer to the references than the RAxML trees. The advantage of MrBayes over RAxML when using the same model of amino acid substitutions can be explained by the influence of the prior distribution of branch lengths, which is implemented in MrBayes. However, this requires confirmation, as MrBayes has plenty of parameters, the testing of which is beyond the scope of this work.

Models for Balanced Minimum Evolution

In the program FastME, several distance-oriented phylogenetic methods are implemented. Also FastME can compute several sequence-to-sequence distance measures to produce a distance matrix, which is the input for all distance methods. We compared these methods with each other using the default distance measure based on the LG substitution model (Le and Gascuel 2008). The results can be seen in supplementary table S8, Supplementary Material online. On all three combined sets, the best average quality was shown by the subtree pruning and regrafting (SPR) search with the Balanced Minimum Evolution (BME) criterion of tree quality.

We used the latter distance method to compare five sequence-to-sequence distance measures: the p-distance, which is the fraction of different letters in two sequences, and likelihood measures based on two substitution models, JTT (the oldest one), and LG (the default one), with and without using gamma distribution of substitution rates. The results are shown in supplementary Table S9, Supplementary Material online.

We observed a slight advantage of the JTT model over the default LG model. More impressive was the observed disadvantage of using the gamma distribution. On 45-sequences alignments, the average distance from trees builded with LG model without gamma was 0.6023 and with gamma was 0.6114, Z-score was 11.8. For 15-sequence alignments even using p-distance led to slightly better results than using the likelihood measures with gamma distribution of rates. The observed phenomenon may be explained with the results of the paper (Guindon and Gascuel 2002). In that work it was shown that, for simulated alignments, the optimal gamma shape parameter for computing sequence-to-sequence distances was much larger than the real one, *i.e.* the parameter used for the simulations. Also, from the results of that work it follows that even infinite value of gamma shape parameter (which means no rate heterogeneity between sites) led to better results of distance methods than the real value, if the evolution was simulated with low deviation from molecular clock (see the right half of Fig. 1 in Guindon and Gascuel 2002).

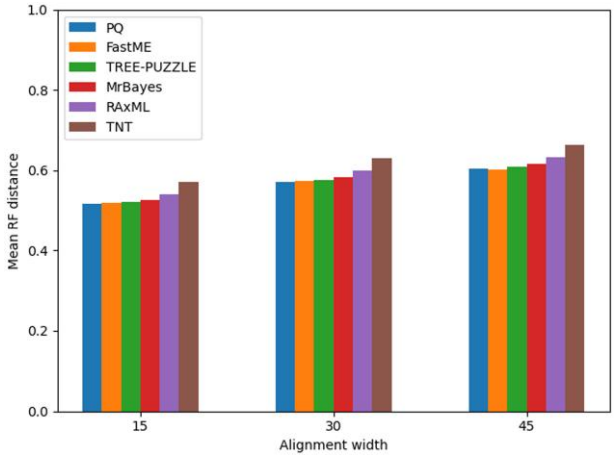

Fig. 3. Bar plots of average RF distances from inferred trees to resolved species trees for the six methods on three combined sets.

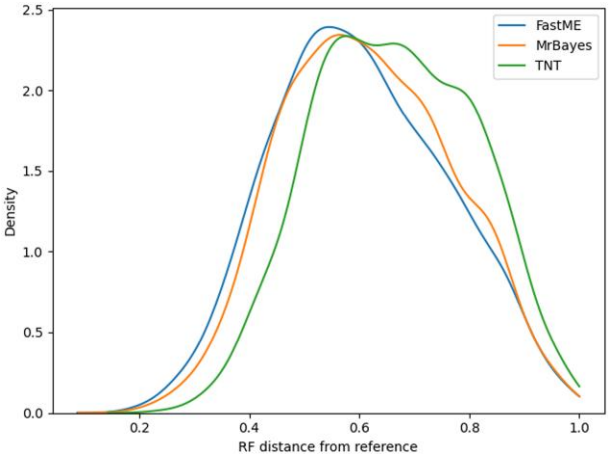

Fig. 4. Density plots of RF distances from trees inferred with three programs to resolved species trees. Input alignments were from the combined set of 45-sequence alignments.

Comparison of Six Methods with Optimized Parameters

We compared six programs: RAxML, MrBayes, FastME, TNT (implementing MP), TREE-PUZZLE (implementing the Quartet Puzzle algorithm), and PQ (implementing the PQ algorithm) on three combined sets. For RAxML, MrBayes, and FastME, we used the parameters that showed best results on our datasets (see previous sections). For RAxML, we used the JTT model with no rate heterogeneity. For MrBayes, we used the ASTRAL consensus of trajectory, with trajectory length set to 1,000 for the 15-sequence set, 2,000 for the 30-sequence set, and 4,000 for the 45-sequence set. For FastME, we used the JTT model without rate heterogeneity for sequence-to-sequence distances, BME criterion, and the SPR search. The parameters for TNT, TREE-PUZZLE, and PQ are described in the Supplementary Material online. The results are listed in supplementary tables S10 and S11, Supplementary Material online; in the supplementary table S10, Supplementary Material online Z-scores were computed for the RF distances to unresolved taxonomic trees and in supplementary table S11, Supplementary Material online for the RF distances to the resolved species trees. Average distances from trees inferred by these programs to the resolved species trees are illustrated in Fig. 3.

Figure 4 illustrates the distributions of the distances to the resolved species trees for the 45-sequence combined set for three most contrasting programs, FastME (demonstrated the best results on our data), TNT (worse results), and MrBayes (medium results).

Distributions in Fig. 4 seem rather similar. The main cause of this is that the distance between inferred and reference trees depends much more on the input alignment than on the inference method. For example, the Pearson correlation between distances from FastME trees to references and from TNT trees to references is about 0.9. At the same time, mean differences between trees inferred with any two of the four methods, TNT, RAxML, MrBayes, and FastME, significantly differ from zero, see supplementary tables S10 and S11, Supplementary Material online. Results of PQ and TREE-PUZZLE are mainly close to the results of FastME.

On three combined sets, FastME, with the selected options, was in average more accurate than MrBayes, RAxML, and TNT; on 45-sequence alignments also than PQ and TREE-PUZZLE. It's noteworthy that even when using p-distances, FastME performed better than RAxML with its best options. Thus, on the 45-sequence combined set, FastME using SPR search, the BME criterion, and p-distance produced a tree closer to the reference than the RAxML tree in 989 cases and outputted a more distant tree in 657 cases, with $Z = -5.87$. However, the superiority of FastME observed on randomly chosen OGs was not observed on some specially selected subsets, see the next subsection.

The superiority of distance methods over ML on natural sequences has long been shown (Krivozubov and Spirin 2010; Gonnet 2012; Penzar et al. 2018). In particular, Gonnet (2012)'s article, being based on a substantial amount of data and published in a popular open access journal, presumably should close the problem. Nevertheless, this article has only been cited 20 times in 11 years (according to Google Scholar), and the majority of researchers continue to consider ML as the most accurate method and distance methods to be of little use for phylogenetic reconstruction. There may be several reasons for this. First, the ML principle appears to be the most theoretically reasonable. Second, there may be some inertia regarding change: many researchers learned about the superiority of ML when first encountering phylogenetic inference. Moreover, on simulated alignments, distance methods usually perform worse than ML and, at times, even worse than MP. The reasons for the obtained differences in quality between ML and distance methods, which use the same evolutionary models for calculating distances, remain unclear.

One possible explanation for the conflicting results on simulated and real alignments is in nonrealistic methods of simulations. This possibility is supported by the opinion of A. Stamatakis, who writes in the RAxML manual (https://sco.h-its.org/exelixis/resource/download/NewManual.pdf, p. 60):

> Q: Why has the performance of RAxML mainly been assessed using real-world data? A: Personal opinion: Despite the unquestionable need for simulated data and trees to verify and test the performance of current ML algorithms the current methods available for generation of simulated alignments are not very realistic. ⟨…⟩ Typically, search algorithms execute significantly less (factor 5–10) topological moves on simulated data until convergence as opposed to real data, *i.e.* the number of successful Nearest Neighbor Interchanges (NNIs) or subtree rearrangements is lower.

However, Guindon and Gascuel (2002) observed that on alignments simulated with certain parameters, namely a high variability of substitution rates among sites and approximate molecular clock, distance methods also outperform maximum likelihood.

We have compared a bayesian method with other methods on natural data for the first time. In the mentioned above article Gonnet (2012), bayesian methods were deliberately left uninvestigated. The author explains this as follows (page 19):

> Bayesian methods for tree building have not been included in this study because they do not follow the PTMS [phylogenetic tree reconstruction from molecular sequences] definition. In principle, a bayesian tree building method produces a probability distribution over all trees given the corresponding priors. If the priors are ignored and only the tree with highest probability is selected, then this is ML, not bayesian. Approaches which build consensus trees from several of the most probable trees produce multifurcating trees which contain less information and hence are not comparable to fully determined trees. Any prior which contains information about the tree which is not extracted from the sequences themselves will violate our assumptions for PTMS.

These reasons are dubious. Firstly, priors are not only for topologies, but also for branch lengths. In particular, in MrBayes the default is an exponential distribution with a mean of 10 as a prior for the tree length (i.e. for the sum of all branch lengths) with partitioning of the tree length to branch lengths according to the Dirichlet distribution with all parameters equal to 1 (Rannala et al. 2012). The *a posterior* probability is calculated taking branch lengths into account and, therefore, its maximum can be achieved at a different topology than the ML, so priors for branch lengths can affect the choice of topology. Secondly, it is possible to build a resolved consensus of an ensemble of trees, for example, with ASTRAL, which does not produce multifurcating trees. On our benchmark the ASTRAL consensus gives, on average, better results compared to ML (see Fig. 2).

Impact of Input Data Properties

Quality of phylogenetic methods may depend on features of input data. To investigate this, for each alignment of the combined sets we have calculated the following properties: alignment length, average p-distance between sequences, average distance from the predicted root to leaves through the RAxML tree (TH), RAxML tree length (TL), and the ratio TL to TH (TL/TH). See details of calculating TH and TL in Materials and Methods. The ratio TL/M should be relatively large in case when most branching events occurred near the root, and relatively small when there are many recent branching close to leaves. See these values in the supplementary file `Combined_alignments_info.xlsx`, Supplementary Material online.

For each property and each of the six programs (PQ, FastME, TREE-PUZZLE, MrBayes, RAxML, TNT), we calculated the Pearson correlations between the property values and the RF distances between inferred trees and references, see supplementary table S12, Supplementary Material online. Alignment length predictably showed negative correlations with the distances in range −0.30 to −0.16 for all six programs. TL and TL/TH demonstrated positive correlations, 0.15 to 0.17 for TL and 0.17 to 0.27 for TL/TH.

To study impact of the properties on relative quality of different programs we have prepared four additional sets of alignments. First two sets consisted of 1,000 most long and 1,000 most short eukaryotic 45-sequence alignments. First set alignments contained 488 to 2,410 columns, second set alignments contained 21 to 64 columns. Mean RF distances to references for trees generated by the six programs on these two sets are illustrated on Fig. 5a, Z-scores for pairwise program comparisons see in supplementary table S13, Supplementary Material online.

Another two prepared sets consisted of 15-sequence alignments from the combined set whose evolution was "shallow branching" for one set and "deeply branching" for another set, see Materials and Methods. The results of the six programs on these sets are illustrated at Fig. 5b and supplementary table S14, Supplementary Material online.

Relative quality of the methods were indeed sufficiently different between sets of alignments with contrasting characteristics. In particular, differences in quality between FastME, MrBayes, and RAxML became neglected on deeply branching alignments, while became more pronounced on shallow branching alignments. For TNT an opposite trend was observed. In particular, TNT performed considerably worse on deep trees, which is consistent with the classical Felsenstein results on parsimony methods (Felsenstein 1978).

The most impressing contrast with the combined sets, composed of OGs randomly selected from the total pool, was the performance of RAxML on long alignments. Namely, RAxML showed the best results on long alignments, outperforming other five programs. These results suggest that dependence of relative quality of methods on features of alignments requires careful consideration. The obvious restrictions of the current version of PhyloBench are the predominance of short protein sequences and that all alignments consist of relatively small numbers of sequences. Relative quality of phylogenetic programs may depend on both dimensions of the input

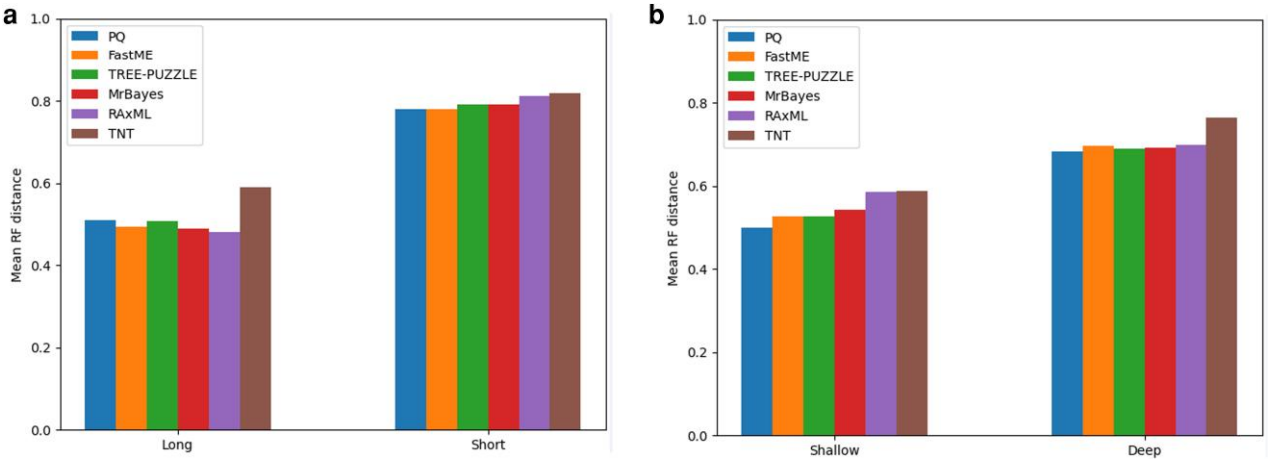

Fig. 5. Bar plots of RF distances from inferred trees to resolved species trees for the six methods on: a) sets of long and short 45-sequence alignments; b) shallow branching and deeply branching sets of 15-sequence alignments.

alignment. In our further work, we intend to overcome these restrictions forming larger sets of OGs and using full-length proteins instead of evolutionary domains.

We also performed comparisons of the six programs separately on three subsets of each combined set, namely on 649 archaeal alignments, 650 bacterial alignments, and 650 eukaryotic alignments, see supplementary tables S15, S16, and S17, Supplementary Material online. The only observation that we derived is that TREE-PUZZLE and PQ performed relatively worse on our eukaryotic alignments, while relative quality of four other methods were approximately the same for all three subsets and for all three alignment sizes.

Online Service

The online service provides three datasets of alignments for download: the 15-sequence, 30-sequence, and 45-sequence combined sets. The user can reconstruct the phylogeny of each alignment from any dataset by the user's method of interest. They can then upload the obtained set of trees to the service and choose one from seven distance measures and one from three variants of reference trees. Distances from the user's trees to the reference trees will be computed and compared with similar distances from the trees built with the six methods, PQ, FastME, TREE-PUZZLE, MrBayes, RAxML, and TNT, all with optimized parameters. The service output is a table containing the numbers of wins, draws, and losses of the user's method compared to the six reference methods, as well as Z-scores for the differences of the distances. The service is available at https://mouse.belozersky.msu.ru/tools/phylobench/. Alignments and reference trees for 36 taxonomic sets are available for download, too.

Conclusion

We have developed PhyloBench, a benchmark for evaluating the quality of phylogenetic programs. Twelve sets, each comprising 60 living species, were selected, and phylogenetic trees of these species were used as references. OGs were generated from the Pfam evolutionary domains present in the proteins of these species. This enabled a comparison of programs by their performance on OGs in regard to distances between the inferred and reference trees. To avoid possible bias related to the topology of the same species tree, we suggest making comparisons not on the 60-sequence alignments, but on alignments of randomly chosen subsets of 15, 30, and 45 sequences from each OG. Our benchmark differs from analogs by two main features: first, it is based on real protein sequences and second, the measure of accuracy of an inferred tree is its distance to a reference tree. The applicability of the benchmark was confirmed by comparisons of phylogenetic inferences based on damaged alignments with inferences based on intact alignments. A number of comparisons of existing phylogenetic tools were performed. In particular, we tested a bayesian program for the first time, and the tests confirmed that distance methods are superior to ML, which, in turn, is superior to MP. On our benchmark, the bayesian program MrBayes showed better quality than ML but worse quality than distance methods. The programs TREE-PUZZLE and PQ performed better compared to MrBayes but, on our 45-sequence alignment dataset, worse compared to the distance method showed the best quality, which is BME. A web-interface providing access to the benchmark is available.

Future Plans

Future plans can be divided into two directions: first, further development of the benchmark itself, and second, research that can be performed using the current version of the benchmark.

It seems natural to create an analogous benchmark for nucleotide alignments. The easiest way would be to take DNA sequences that encode proteins from PhyloBench. However, it would be not an optimal solution, because these proteins are too distant from each other. In our

opinion, a nucleotide phylogenetic benchmark should consist of either noncoding RNA sequences or DNA segments under neutral evolution (such as intergenic transcribed spacers) from closely related species. One more possibility is to use genes of almost identical proteins of closely related species, in this case, the phylogenetic signal would be in synonymous sites, which are also under (almost) no selection pressure. For all these variants selection of an appropriate number of alignments will be significantly more complicated than for proteins.

Another way of future development of the benchmark is using full-length protein sequences instead of sequences of protein domains. Interdomain regions may show different evolutionary patterns compared to domains. We have performed an attempt to use such sequences in Sigorskikh et al. (2022) and we will continue this work.

A number of studies can be conducted based even on the current version of the benchmark. First of all, the dependence of the relative quality of phylogenetic methods on different features of the input alignment should be investigated in much more details.

Another promising direction could be to investigate the quality of consensus making programs, such as ASTRAL and its analogs, in comparison with each other and with using supermatrices.

Materials and Methods

Algorithm for Selecting Orthologous Groups

Twelve sets of 60 living species each were selected. A detailed description of the selection procedure is in the supplementary text, Supplementary Material online. The lists of species are in the supplementary file, `Species.xlsx`, Supplementary Material online.

OGs were composed for each species set from each Pfam family that was represented by protein domains in every species of the set. Each composed OG contained exactly one protein domain sequence from each species. OGs were composed with the following simplified procedure. Fix a Pfam family and a set of species such that in each of these species there is a protein with a domain from the chosen Pfam family. then:

- Select a species *M* with the smallest number of sequences from the Pfam family.
- For each sequence *x* from *M*:
  - found best bidirectional hits (BBHs) in other species;
  - if there are BBHs for *x* in all other species from the set, then form an OG consisting of these BBHs together with *x*, else skip *x*.

BBHs were found using the program blastp from the BLAST+ (Altschul et al. 1997) package version 2.7.1+, with the following parameters: word size 2, window size 0 (which means 1-hit algorithm), word score threshold 7, switch off compositional based statistics and all filters, other parameters by default. A sequence *x* from *M* and a sequence *y* from another species *N* were regarded forming a BBH, if the alignment score of *x* with *y* is not less than all other alignment scores of *x* with sequences from *N* or of *y* with sequences from *M*. OGs that contained less than 45 pairwise different sequences were removed.

Alignments

From each OG, three subsets of 15, 30, and 45 pairwise different sequences were randomly and independently selected. Each subset was aligned using MUSCLE v3.8.1551 (Edgar 2004) with the default settings. Alignments with less than ten non-conserved columns without gaps were removed. Thus, for each species set, we constructed a set of OGs and three "taxonomic" sets of alignments with 15, 30, and 45 sequences in each alignment.

Combined Sets of Alignments

The combined sets of alignments were based on all 649 archaeal OGs, 650 randomly chosen bacterial OGs and 650 randomly chosen eukaryotic OGs (see details of the random choice in the supplementary text, Supplementary Material online). The corresponding 15, 30, and 45 sequence alignments formed three combined sets of 1,949 alignments each.

Phylogenetic Programs

In this study, we used the following phylogenetic programs: TNT (Goloboff and Catalano 2016), RAxML (Stamatakis 2014), MrBayes (Ronquist et al. 2012), PQ (Penzar et al. 2018), TREE-PUZZLE (Schmidt et al. 2002), FastME (Lefort et al. 2015), and IQ-Tree (Nguyen et al. 2015). A description of the used versions and options are in the supplementary text, Supplementary Material online. We intentionally limited ourselves to comparing only those methods that produced fully resolved trees. To obtain resolved trees in all cases, we use ASTRAL (Zhang et al. 2018) to make consensus trees from ensembles of trees outputted by TREE-PUZZLE and MrBayes. Recognizing that unresolved trees are more sensible than fully resolved trees if the original data is insufficient to resolve some nodes, we intend to investigate this issue in future work.

Species Trees

The species tree for each species set was obtained from the NCBI Taxonomy database (Federhen 2012) of the 2021 March 12 version. These trees were not completely resolved (*i.e.* not binary or some nodes had more than two descendants). A completely resolved tree of 60 taxa should contain 57 inner branches, while the obtained trees contain 16 to 43 inner branches, see supplementary table S3, Supplementary Material online. To form binary trees, additional branches were obtained as follows. For each OG, all 60 its sequences were aligned. If the obtained alignment contained 10 or more nonconserved columns without gaps, then three trees were inferred from this alignment with TNT, RAxML, and FastME. For this task, we took the default options of these programs. Since RAxML does not provide a default model, we chose the model

PROTGAMMAAUTO, which seemed to be the most natural choice. For each species set, the completely resolved species tree was the ASTRAL consensus of all inferred trees of the set, with constraints derived from NCBI Taxonomy.

Tree-to-tree Distance Measures

The following seven measures were tested: Robinson–Foulds (RF) distance (Robinson and Foulds 1981), a modified Robinson–Foulds distance (MRF), L_1-distance (Williams and Clifford 1971), L_2-distance (Penny et al. 1982), Quartet (Q) distance (Estabrook et al. 1985), Agreement (A) distance (Gordon 1980), and Agreement L-distance (AL) (Goddard et al. 1994). A short description of these measures is in the supplementary text, Supplementary Material online.

Comparison of Phylogenetic Methods

Each pairwise comparison was performed as follows. Given a set of n multiple sequence alignments (*e.g.* one of the combined sets), two trees were built from each alignment with two methods being compared. Let s_i be the distance from the reference tree to the tree built from the ith alignment with one method, r_i be the distance to the tree built from the same alignment with another method, and SE be the standard error of the set of differences $\{s_i - r_i\}$. The main measure for the comparison of the methods is then:

$$Z = \frac{\sum_i (s_i - r_i)/n}{SE} \quad (1)$$

i.e. the Z-score for the average difference between two distances. Note that if n is large enough, then the Z-score under the null hypothesis (that is the same accuracy of the two methods) is distributed according to the standard normal distribution, thus Z-scores can be easily converted to p-values.

Damaged Alignments

To test reference trees and distance measures, the following procedure was applied to each taxonomic alignment set:

1) each alignment was damaged by replacing one-fifth of columns with columns containing the letter "A" in all sequences;
2) the phylogeny was inferred with the program TNT from both intact and damaged alignments;
3) all distance measures from the two obtained trees to the species tree were computed.

The procedure was repeated five times, first replacing the 1st, 6th, 11th, etc. columns of each alignment, then the 2nd, 7th, 12th, *etc*.

Comparison of Tree-to-tree Distance Measures

With each measure we compared phylogenetic inference based on intact alignments against one based on damaged alignments, with the program TNT with default options. For each of five variants of damage the Z-score was computed with the formula (1), where reconstructions from the intact alignments, from one hand, and from the damaged alignments, from the other hand, stay for two compared methods. Next, the median of these five Z-scores was computed.

The just described procedure was performed with r_i and s_i computed with all seven tree-to-tree distance measures of distances between the inferred and the resolved species tree, *e.g.* the ASTRAL consensus of inferred trees with taxonomic constraints. Additionally, in the same manner we tested RF, MRF, and Q measures of distances to the unresolved taxonomic trees. The distance measure providing the largest median Z-score for most taxonomic sets was regarded as the best. This procedure allowed us to choose the measure that is the most sensitive to errors in phylogenetic reconstruction.

Simultaneously, this procedure validated the applicability of species trees for benchmarking phylogenetic methods. A species tree was regarded as satisfactory, if these Z-scores were large enough (all greater than 1.65, which means confidence level of 0.05).

Calculation of Alignment Properties

IQ-Tree outputs several characteristics for each input alignment, including its length (*i.e.* the number of columns) and other characteristics such as the number of distinct patterns, parsimony-informative sites, singleton sites, and constant sites. Also, we computed numbers of columns with gaps, the average p-distance between sequences and three characteristics derived from the RAxML tree inferred from the alignment. RAxML was used with JTT model and no rate heterogeneity. The three characteristics were tree height (TH), tree length (TL), and their ratio (TL/TH). To compute TH each tree was rooted with the option "`-f I`" of RAxML (see RAxML Manual for details). For each leaf, its distance to the root was computed as the sum of lengths of branches composing the path from the root to the leaf. TH is the mean value of the obtained root-to-length distances. TL was computed as the total sum of tree branches. Properties of all alignments of the combined sets are available in the supplementary file `Combined_alignments_info.xlsx`, Supplementary Material online.

Selection of Deeply Branching and Shallow Branching Sets

From each 60-sequence family of the combined set, two 15-sequence subsets were selected. The first one called "deeply branching" was composed from 15 sequences most distant (according to p-distance) from each other. The second one called "shallow branching" consists of three most distant sequences and 12 sequences most close to any of the three.

Supplementary Material

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

Authors' Contributions

S.S. planned the work, S.S., A.S., A.E., and D.P. wrote the software, S.S., A.S., and A.E. performed the computational experiments and prepared the figures, S.S., A.S., and A.K. analysed the results, S.S. and A.K. wrote the manuscript.

Authors' Information

Sergey Spirin is a leading researcher in Belozersky Institute of Physico-Chemical Biology of Lomonosov Moscow State University. His research interests include bioinformatics, computational phylogenetics and computational structural biology.

Andrey Sigorskikh is a PhD student at the Faculty of Bioengineering and Bioinformatics at Lomonosov Moscow State University. His research areas are computational phylogenetics and machine learning in biology.

Aleksei Efremov is a PhD student at the Faculty of Bioengineering and Bioinformatics at Lomonosov Moscow State University. His research areas are phylogenetics, machine learning in biology and bioinformatics.

Dmitry Penzar is an assistant professor in the Faculty of Bioengineering and Bioinformatics at Lomonosov Moscow State University and a researcher in Artificial Intelligence Research Institute. His research interests are in areas of computational biology, including machine learning.

Anna Karyagina is a professor and chief scientist in Gamaleya National Research Center for Epidemiology and Microbiology, Institute of Agricultural Biotechnology, and Belozersky Institute of Physico-Chemical Biology of Lomonosov Moscow State University with a focus on molecular biology and genome analysis.

Funding

The work was supported by the Russian Science Foundation [grant number 21-14-00135 to S.S., A.S., A.E. and A.K].

Conflict of interest statement. None declared.

Data Availability

The taxonomic sets of alignments and reference trees are downloadable from https://mouse.belozersky.msu.ru/tools/phylobench/data. The combined sets are available at https://mouse.belozersky.msu.ru/tools/phylobench/. The source code (in ANSI C) for the programs implementing tree-to-tree distance measures is available at https://github.com/belozersky321/UMAST (Agreement distance and Agreement L-distance) and at https://github.com/belozersky321/TreeDist (other measures).

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997:**25**(17):3389–3402. https://doi.org/10.1093/nar/25.17.3389.

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004:**32**(5):1792–1797. https://doi.org/10.1093/nar/gkh340.

Estabrook GF, McMorris FR, Meacham CA. Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. *Syst Biol*. 1985:**34**(2):193–200. https://doi.org/10.2307/sysbio/34.2.193.

Federhen S. The NCBI Taxonomy database. *Nucleic Acids Res*. 2012:**40**(D1):D136–D143. https://doi.org/10.1093/nar/gkr1178.

Felsenstein J. Cases in which parsimony or compatibility methods will be positively misleading. *Syst Zool*. 1978:**27**(4):401–410. https://doi.org/10.2307/2412923.

Goddard W, Kubicka E, Kubicki G, McMorris FR. The agreement metric for labeled binary trees. *Math Biosci*. 1994:**123**(2):215–226. https://doi.org/10.1016/0025-5564(94)90012-4.

Goloboff P, Catalano S. TNT, version 1.5, including a full implementation of phylogenetic morphometrics. *Cladistics*. 2016:**32**(3):221–238. https://doi.org/10.1111/cla.2016.32.issue-3.

Gonnet GH. Surprising results on phylogenetic tree building methods based on molecular sequences. *BMC Bioinformatics*. 2012:**13**(1):148. https://doi.org/10.1186/1471-2105-13-148.

Gordon AD. On the assessment and comparison of classifications. In Tomassine R., editor, *Analyse de données et informatique*, pages 149–160. Le Chesnay, France: INRIA. 1980.

Gu X, Fu Y-X, Li W-H. Maximum likelihood estimation of the heterogeneity of substitution rates among nucleotide sites. *Mol Biol Evol*. 1995:**12**(4):546–557. https://doi.org/10.1093/oxfordjournals.molbev.a040235.

Guindon S, Gascuel O. Efficient biased estimation of evolutionary distances when substitution rates vary across sites. *Mol Biol Evol*. 2002:**19**(4):534–543. https://doi.org/10.1093/oxfordjournals.molbev.a004109.

Hollich V, Milchert L, Arvestad L, Sonnhammer ELL. Assessment of protein distance measures and tree-building methods for phylogenetic tree reconstruction. *Mol Biol Evol*. 2005:**22**(11):2257–2264. https://doi.org/10.1093/molbev/msi224.

Jones D, Taylor W, Thornton J. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci*. 1992:**8**:275–282. https://doi.org/10.1093/bioinformatics/8.3.275.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*. 2017:**14**(6):587–589. https://doi.org/10.1038/nmeth.4285.

Krivozubov MS, Spirin SA. Comparison of protein phylogeny reconstruction methods using natural protein sequences. *Moscow Univ Biol Sci Bull*. 2010:**65**(4):139–141. https://doi.org/10.3103/S0096392510040036.

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.

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, et al. Pfam: the protein families database in 2021. *Nucleic Acids Res*. 2021:**49**(D1):D412–D419. https://doi.org/10.1093/nar/gkaa913.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 2015:**32**(1):268–274. https://doi.org/10.1093/molbev/msu300.

Penny D, Foulds LR, Hendy MD. Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences. *Nature*. 1982:**297**(5863):197–200. https://doi.org/10.1038/297197a0.

Penzar D, Krivozubov M, Spirin S. PQ, a new program for phylogeny reconstruction. *BMC Bioinformatics*. 2018:**19**(1):374. https://doi.org/10.1186/s12859-018-2399-4

Rannala B, Zhu T, Yang Z. Tail paradox, partial identifiability, and influential priors in Bayesian branch length inference. *Mol Biol Evol*. 2012:**29**(1):325–335. https://doi.org/10.1093/molbev/msr210.

Robinson DR, Foulds LR. Comparison of phylogenetic trees. *Math Biosci*. 1981:**53**(1–2):131–147. https://doi.org/10.1016/0025-5564(81)90043-2.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. MRBAYES 3.2: efficient Bayesian phylogenetic inference and model selection across a large model space. *Syst Biol*. 2012:**61**(3):539–542. https://doi.org/10.1093/sysbio/sys029.

Schmidt HA, Strimmer K, Vingron M, von Haeseler A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*. 2002:**18**(3):502–504. https://doi.org/10.1093/bioinformatics/18.3.502.

Sigorskikh AI, Latortseva DD, Karyagina AS, Spirin SA. How often does filtering of alignment columns improve the phylogenetic inference of two-domain proteins? *Biochemistry (Moscow)*. 2022:**87**(12-13):1689–1698. https://doi.org/10.1134/S0006297922120239.

Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014:**30**(9):1312–1313. https://doi.org/10.1093/bioinformatics/btu033.

Whelan S, Goldman N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol*. 2001:**18**(5):691–699. https://doi.org/10.1093/oxfordjournals.molbev.a003851.

Williams WT, Clifford HT. On the comparison of two classifications of the same set of elements. *Taxon*. 1971:**20**(4):519–522. https://doi.org/10.2307/1218253.

Wu M, Chatterji S, Eisen JA. Accounting for alignment uncertainty in phylogenomics. *PLoS One*. 2012:**7**(1):e30288. https://doi.org/10.1371/journal.pone.0030288.

Zhang C, Rabiee M, Sayyari E, Mirarab S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*. 2018:**19**(S6):153. https://doi.org/10.1186/s12859-018-2129-y.

Zhou X, Shen X, Hittinger CT, Rokas A. Evaluating fast maximum likelihood-based phylogenetic programs using empirical phylogenomic data sets. *Mol Biol Evol*. 2017:**35**(2):486–503. https://doi.org/10.1093/molbev/msx302.