PhyloBench, a benchmark for evaluation of phylogenetic programs

Andrey Sigorskih [1], Alexey Efremov [1], Dmitry Penzar [1], Anna Karyagina [2, 3, 4], Sergey Spirin[2, 5, 6*]

[1] *Faculty of Bioengineering and Bioinformatics of Lomonosov Moscow State University, Moscow, Russia*
[2] *Belozersky Institute, Lomonosov Moscow State University, Moscow, Russia*
[3] *Gamaleya National Research Center of Epidemiology and Microbiology, Moscow, Russia*
[4] *All-Russia Research Institute of Agricultural Biotechnology, Moscow, Russia*
[5] *Scientific Research Institute for System Analysis of the Russian Academy of Sciences, Moscow, Russia*
[6] *National Research University Higher School of Economics, Moscow, Russia*
* sas@belozersky.msu.ru

The work was supported by the Russian Science Foundation grant 21-14-00135

Summary

Research Objective. Most published comparisons of phylogenetic methods either are based on simulated alignments or use some calculated features, as log likelihood, for the evaluation. The aim of the present work is a benchmark that allows such comparison on large sets of natural orthologous protein sequences using species trees as reference trees.

Materials and Methods. PhyloBench consists of protein sequence alignments and of reference trees for these alignments. We used sequences of evolutionary protein domains extracted from the Pfam database. For 12 sets of 60 living species each, representing all major taxa of cellular organisms, we formed as many as possible orthologous groups of domains from proteins of those 60 species. The species tree was constructed starting with NCBI Taxonomy, unresolved nodes of the tree were resolved using branches of the trees inferred from all obtained orthologous groups. From each 60-sequence alignment we extract three subalignments, of 15, 30, and 45 sequences. These subalignments form the benchmark. For testing the benchmark we used comparison of inferences made with real sequence alignments of domains and those made with artificially damaged alignments. For comparison of inferred trees with reference trees we used a number of tree comparison measures and chose the measure that allows us to obtain the maximum statistical significance during the test.

Results. The Robinson–Foulds distance is proved to be the best tree comparison measure. We demonstrated a statistically significant difference between results obtained from real and damaged alignments thereby confirming applicability of our benchmark. We created combined set of 1949 alignments equally representing Archaea, Bacteria, and Eukaryotes. Using the combined set, we performed a number of comparisons of phylogenetic methods and their parameters. In particular, we confirmed recent results that alignment filtering does not improve the accuracy of phylogenetic inference and that distance methods, such as minimum evolution, are superior to maximum likelihood and maximum parsimony.

Conclusion. PhyloBench allows to evaluate a quality of any tool that infers phylogeny from a protein sequence alignment.

Density plot of distances from inferred to reference trees

We tested the following phylogenetic programs:
•TNT, an implementation of maximum parsimony method
•RAxML, an implementation of maximum likelihood method
•MrBayes implementing Monte-Carlo Markov chain (MCMC) search in tree space with *a posterior* probability of a tree as the objective function (so-called bayesian phylogenetic inference)
•PQ, an implementation of our original algorithm
•TREE-PUZZLE, an implementation of quartet-puzzling method
•FastME implementing a number of distance-oriented methods

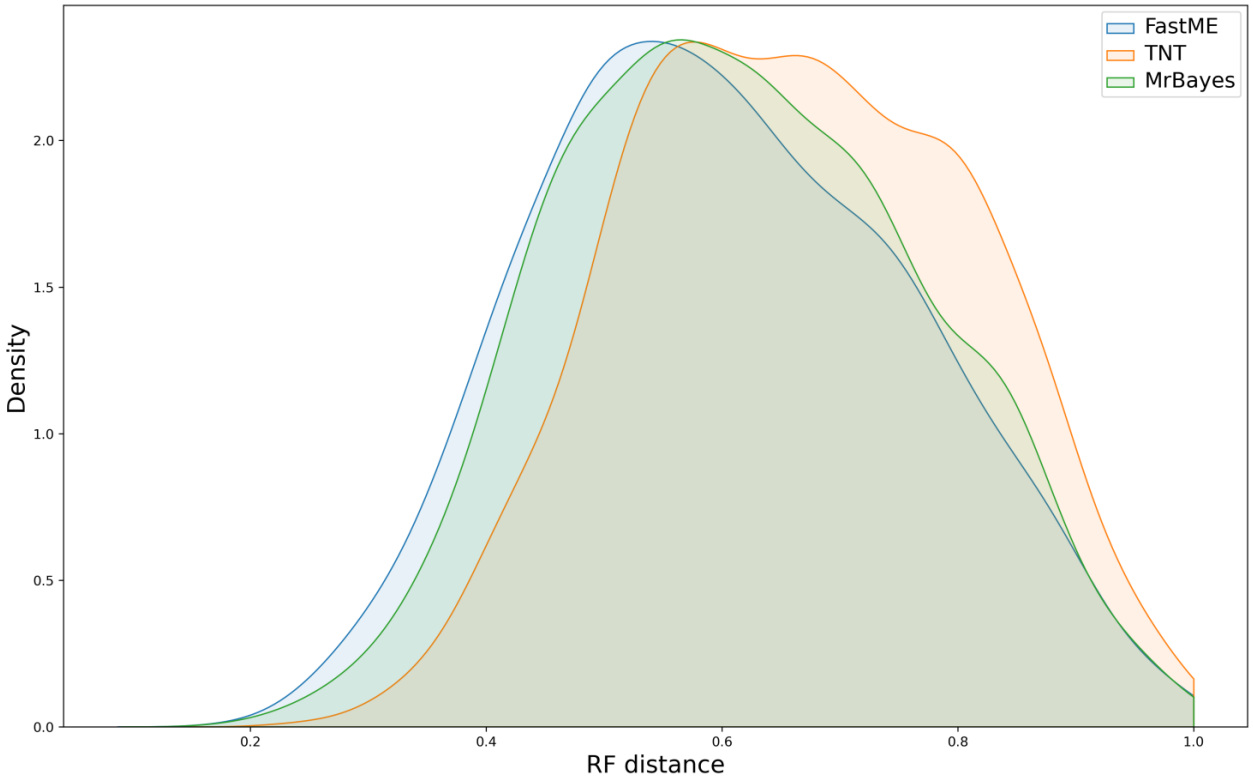
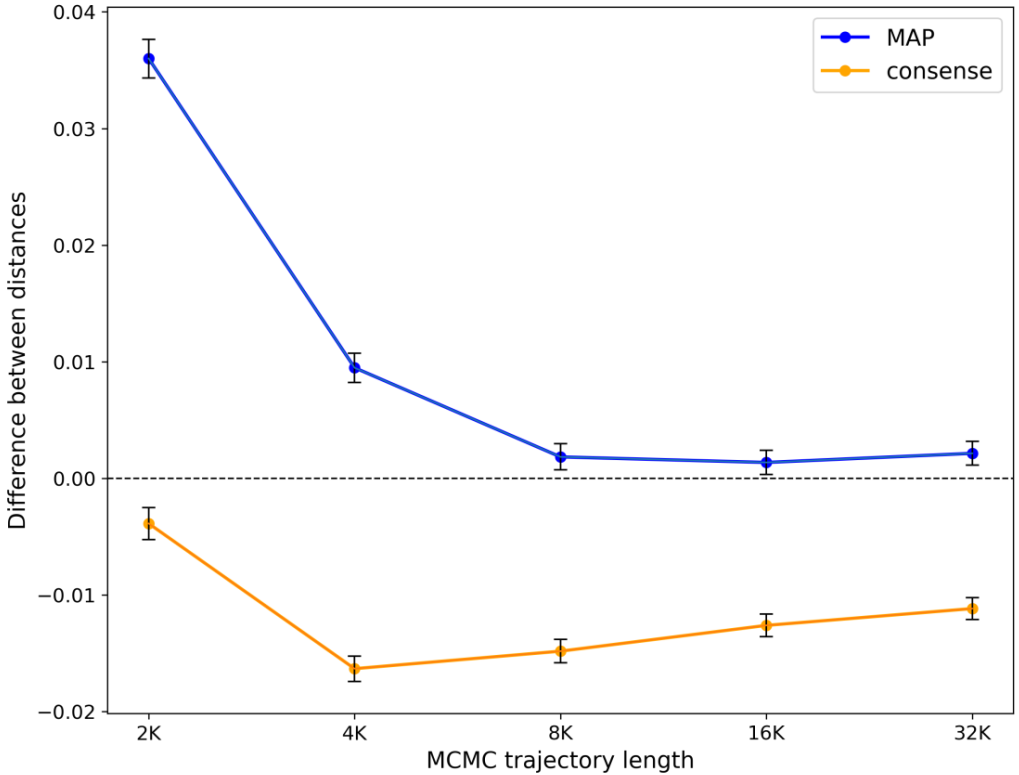


Figure 1. Density plot of RF-distances from inferred trees to reference trees for three methods on the combined set of 45-sequence alignments. For MrBayes, the Jones substitution model, 4000 steps of MCMC and consensus of the outputted ensemble was used. For FastME, the results of balanced minimum evolution method is shown.

Dependence of MrBayes results on the number of MCMC steps

Figure 2. Dependence of average quality of phylogenetic inference by MrBayes on MCMC trajectory length, in comparison with RAxML. The vertical axis corresponds to the average difference between Robinson–Foulds distances from MrBayes and RAxML trees to reference trees. Error bars reflect standard errors of the average differences. Calculations were performed on the combined sets of alignments of 45 sequences. The blue line is for trees with maximum a posterior probabilities (MAP), the yellow line is for consensuses on MCMC trajectories.

MAP trees were, on average, more distant from references than consensus trees for any number of iterations. For the consensus trees, there was an optimal number of iterations.



Comparison of six methods

Table. Comparison of six methods on the combined set of 45-sequence alignments. In columns 2–7 there are Z-scores for pairwise comparisons, a negative number means the superiority of the method in row. Column 8 contains the mean RF distances from the references, and column 9 contains the calculation times time relative to the time of TNT.
For all programs the best sets of parameters were used. In particular, FastME is used with subtree pruning and regrafting (SPR) search for the best tree according to balanced minimum evolution criterion. This method is proved to be the best of all methods tested.

	TNT	RAxML	MrBayes	TREE-PUZZLE	PQ	FastME	Mean RF	Time
TNT		20.7	32.3	34.2	34.1	39.8	0.6614	1
RAxML	−20.7		15.1	15.0	15.2	21.3	0.6276	444.9
MrBayes	−32.3	−15.1		4.50	5.09	10.3	0.6134	80.0
TREE-PUZZLE	−34.2	−15.0	−4.50		0.95	5.75	0.6058	318.3
PQ	−34.1	−15.2	−5.09	−0.95		4.75	0.6047	895.0
FastME	−39.8	−21.3	−10.3	−5.75	−4.75		0.5993	0.66

The pairwise comparisons were performed as follows: given a set of *1949* multiple sequence alignments, two trees were built from each alignment with two compared methods. Let s_i be the distance from the reference tree to the *i*th tree built with one method, r_i be the same distance to the *i*th tree built 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: $Z = \frac{\sum_i (s_i - r_i)/n}{\mathrm{SE}}$, i.e., the Z-score for the average difference between two distances.

Key points

- PhyloBench is a benchmark for evaluating the quality of phylogenetic programs. It is based on natural, orthologous protein sequences. The measure of accuracy of an inferred tree is its distance to the species tree.
- A number of tree-to-tree distance measures were tested, and the most reliable results were obtained using the Robinson–Foulds distance.
- A number of popular phylogenetic programs were tested, and the most accurate was shown to be balanced minimum evolution implemented in the FastME program.
- Bayesian phylogenetic inference, if used with consensus of the MCMC trajectory, is more accurate than maximum likelihood but less accurate than distance methods.
- Alignments and reference trees of the benchmark are available online at https://mouse.belozersky.msu.ru/phylobench/pb.html together with a web-interface allowing for the semi-automatic comparison of a user's method with a number of popular programs.