

PhyloBench, a benchmark for evaluation of phylogenetic programs

Sigorskih A.^{1*}, Efremov A.¹, Penzar D.¹, Karyagina A.^{2, 3, 4}, Spirin S.^{2, 5, 6*}

¹ Faculty of Bioengineering and Bioinformatics of Lomonosov Moscow State University, Moscow, Russia

² Belozersky Institute of 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, RAS, Moscow, Russia

⁶ National Research University Higher School of Economics, Moscow, Russia

* sas@belozersky.msu.ru

Key words: phylogenetic inference, benchmark

Motivation and Aim: There are a lot of algorithms and programs for phylogenetic inference based on a multiple sequence alignment. Many programs allow users to choose a number of parameters and options. Different programs often produce different results, the same holds for different options of a same program. 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 creation of a benchmark that allows such comparison on large sets of natural orthologous protein sequences using species trees as reference trees.

Results: PhyloBench consists of protein sequence alignments and of reference trees for these alignments. We used sequences of evolutionary protein domains extracted from the Pfam [1] database. To select the sequences, we first fixed a taxonomic group of species and selected 60 species from this group. Then using home-made scripts 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. This procedure was repeated 12 times, for different taxonomic groups covering all major taxa of cellular organisms. As a result, the benchmark consists of 12 taxonomic sets, for each of them there are sets of alignments of 15, 30, and 45 sequences randomly selected from 60-sequence orthologous groups. Also PhyloBench includes three “Combined” sets containing 15-, 30-, and 45-sequence alignments selected from taxonomic sets and representing three domains of life (Archaea, Bacteria and Eukaryota) in equal proportions. 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. As a result, the Robinson–Foulds distance [2] is proved to be the best tree comparison measure. We demonstrate a statistically significant difference between results obtained from real and damaged alignments thereby confirming applicability of our benchmark. Using PhyloBench, 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 [3] and that distance methods, such as minimum evolution, are superior to maximum likelihood and maximum parsimony [4].

Availability: The benchmark, i.e. alignments and reference trees, is available online <http://mouse.belozersky.msu.ru/phylobench/pb.html> together with the service that allows to compare trees inferred by any user's method with trees inferred by three widely used methods: maximum parsimony, maximum likelihood and BioNJ.

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

Acknowledgements: The work was supported by the Russian Science Foundation grant No. 21-14-00135.

References

1. Mistry J. et al. Pfam: The protein families database in 2021. *Nucleic Acids Res.* 2021;49(D1):D412-D419.
2. Robinson D.R., Foulds L.R. Comparison of phylogenetic trees. *Math Biosci.* 1981;53(1-2):131-147.
3. Tan G. et al. Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference. *Syst Biol.* 2015;64(5):778-791.
4. Gonnet G.H. Surprising results on phylogenetic tree building methods based on molecular sequences. *BMC Bioinformatics.* 2012;13:148.