A benchmark for evaluation of phylogeny reconstruction programs

Sergei Spirin
Belozersky Institute of Physico-Chemical Biology, Lomonosov Moscow State University, Moscow, Russia
sas@belozersky.msu.ru

Introduction

There are a lot of algorithms and programs for reconstruction of phylogeny of a set of proteins basing on multiple sequence alignment. Many programs allow users to choose a number of parameters, for example, a model for maximum likelihood programs.

Different programs and different parameters often produce different results. However at the moment all published benchmarks for evaluation of relative accuracy of programs or different choices of parameters are based on simulated sequences.

The aim of the present work is to create a benchmark that allows a comparison of phylogenetic programs on large sets of alignments of natural protein sequences.

Methods

We used orthologous series extracted from Pfam [1] families of evolutionary protein domains. The benchmark consists of sequence alignments of domains. Each alignment includes one sequence per species, thus a phylogenetic reconstruction based on an alignment can be compared with the phylogeny of corresponding species.

The benchmark consists of a number of subdivisions created as follows.

1. Select a taxon. Three taxons were used: Fungi, Metazoa, and Proteobacteria.

2. Select a set of species from the selected taxon. The following species sets were used: 45 fungal species from different genera, 45 proteobacterial species from different families and 25 metazoan species. Fungal and proteobacterial species were selected by a special script maximizing the total number of Pfam families presented in all these species. Metazoan species were selected manually to obtain a set whose phylogeny is fully resolved and unambigous according to the current knowledge in animal systematics.

3. Select all Pfam families presented in all selected species.

4. From each selected Pfam family extract all orthologous series representing all selected species. For the current version of the benchmark, the selection of orthologous series was performed with the following simplified procedure:
– select a species M with the smallest number of sequences from the Pfam family,
– for each sequence X from M found best bidirectional hits (BBH) in other species,
– if there are BBH from all species, then form an orthologous series from these BBH together with X, otherwise skip X.

5. Select a number n not greater than the number of species. At the moment, $n = 10, 15, 25$ for Metazoa and $n = 15, 30, 45$ for Fungi and Proteobacteria are used. From each orthologous series randomly select n sequences.

6. Estimate the species tree. For the Metazoa set, we use the tree known from the zoological systematics. For other sets, the estimations of the species trees are consensuses of trees reconstructed from all full-sized (i.e., 45-species) orthologous series by a number of phylogenetic programs. These estimated species trees can contain some errors but (as we demonstrated) are close enough to the real trees.

A comparison of two methods (two phylogenetic programs or two sets of parameters) is performed as follows.

1. Choose a subdivision of the benchmark, i.e., a set of orthologous series (OS) of one size from one taxon.

2. For each OS:
   – reconstruct two phylogenetic trees with two methods;
   – obtain the species tree by restriction of the entire organism tree to the set of species presented in the OS;
   – measure a number of tree-to-tree distances from the reconstructed trees to the species tree.

   If one of the reconstructed trees is closer to the etalon tree according to all distance measures, then we say that on this particular OS one reconstruction method works better than another one.

3. Calculate numbers of OS for which each method works better than another one. Compare two numbers with the sign test to obtain p-value. If the p-value is less than 0.001, then we can conclude that one method is more accurate (on this particular set of OS) than another method.

In the point 2, we use two tree-to-tree distance measures. One is well-known Robinson – Foulds distance [2], that is the number of different branches (splits) in two trees. Another is so-called combinatorial distance, described in [3].

To evaluate our benchmark, we aligned sequences of all OS of all sets with Muscle [4] and compared two "methods". The first is program TNT [5] with default parameters. The second is the same program, but using alignments with first one-fourth of columns being removed. Obviously the second "method" should be less accurate than the first one.

Results

The Muscle alignments of orthologous series and organism trees are available at `http://mouse.genebee.msu.ru/phylobench` .

Comparison of trees reconstructed from entire alignments with the trees reconstructed from shortened alignments in all cases shows that the first "method" is more accurate, with the p-value less than 0.001. It demonstrates that all subdivisions of the benchmark are suitable for comparison of methods. In particular, it means that possible errors in organism trees and in selecting orthologous series are at an acceptable level.

References

1. Finn R.D., Coggill P., Eberhardt R.Y., et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016. 44(D1):D279-D285. DOI: 10.1093/nar/gkv1344

2. Robinson D.R., Foulds L.R. Comparison of phylogenetic trees. Mathematical Biosciences. 1981. 53(1–2):131–147. DOI: 10.1016/0025-5564(81)90043-2

3. Krivozubov M., Goebels F., Spirin S. Estimation of relative effectiveness of phylogenetic programs by machine learning. J Bioinf Comp Biol. 2014. 12(2):1441004. DOI: 10.1142/S0219720014410042

4. Edgar R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004. 32(5):1792–1797. DOI: 10.1093/nar/gkh340

5. Goloboff P.A., Farris J.S.„ Nixon K.C. TNT, a free program for phylogenetic analysis. Cladistics. 2008. 24(5):774-786. DOI: 10.1111/j.1096-0031.2008.00217.x