A benchmark for evaluation of phylogeny reconstruction programs

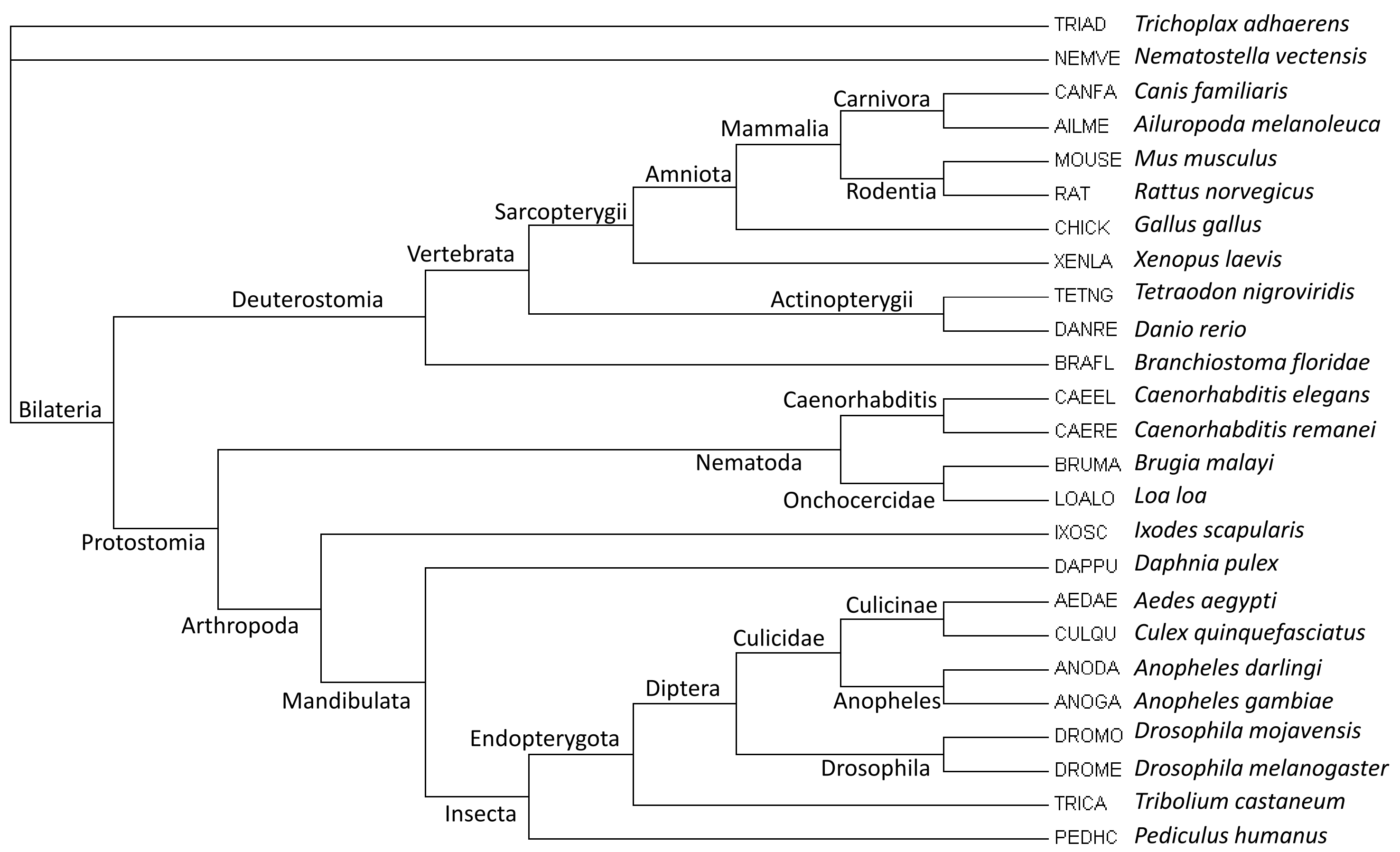
Sergei Spirin sas@belozersky.msu.ru

Belozersky Institute of Physico-Chemical Biology, Moscow, Russia

Introduction

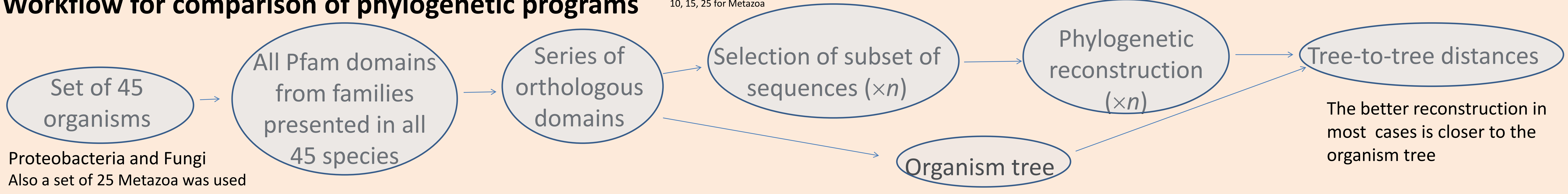
There are a lot of algorithms and programs for reconstruction of phylogeny of a set of proteins basing on multiple sequence alignment. Many programs allow users to choose a number of parameters, for example, a model for maximum likelihood programs. Different programs and different parameters often produce different results. However at the moment there are no published benchmarks for evaluation of relative accuracy of programs or different choices of parameters on natural (not simulated) data.

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



Organism tree of 25 metazoans

Workflow for comparison of phylogenetic programs



Subsets of 15, 30, 45 sequences for Fungi and Proteobacteria, 10, 15, 25 for Metazoa

Set of 45 organisms

Proteobacteria and Fungi
Also a set of 25 Metazoa was used

All Pfam domains from families presented in all 45 species

Series of orthologous domains

Selection of subset of sequences (×*n*)

Phylogenetic reconstruction (×*n*)

Tree-to-tree distances

Organism tree

The better reconstruction in most cases is closer to the organism tree

Testing

To evaluate our benchmark, we compared two "methods".
The first is program TNT with default parameters. The second is the same program, but using alignments of orthologous series (OS) with first one-fourth of columns being removed.
Obviously the second "method" should be less accurate than the first one.

Dataset	# OS	1st is better	2nd is better	P-value
Metazoa-10	1499	337	215	$2.3\cdot10^{-7}$
Metazoa-15	1283	366	236	$1.3\cdot10^{-7}$
Fungi-15	1191	401	247	$1.6\cdot10^{-9}$
Proteobacteria-15	784	253	121	$7.7\cdot10^{-12}$
Metazoa-25	969	271	214	0.011
Fungi-30	1004	395	162	$1.9\cdot10^{-23}$
Proteobacteria-30	783	293	101	$8.6\cdot10^{-23}$
Fungi-45	827	346	153	$3.2\cdot10^{-18}$
Proteobacteria-45	780	281	109	$1.3\cdot10^{-18}$

Availability

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

Usage case

The tables contain results of comparing three programs: TNT[1] (based on maximum parsimony principle), RAxML[2] (based on Maximum Likelihood principle) and FastME[3] (based on Minimum Evolution principle) on two datasets.
In cells: numerator is number of OS for which the left method is better, denominator for which the top one is better.

Fungi-30

	FastME	RAxML	TNT
FastME		483/166	642/64
RAxML	166/483		440/196
TNT	64/642	196/440	

Proteobacteria-30

	FastME	RAxML	TNT
FastME		335/88	440/55
RAxML	88/335		306/126
TNT	55/440	126/306	

[1] The program TNT is being made available with the sponsorship of the Willi Hennig Society, and is published in: P. Goloboff, J. Farris, and C. Nixon. Cladistics 24, 2008.
[2] The program RAxML is published in: A. Stamatakis. Bioinformatics 30(9), 2014. The version 8.2.8 with the model PROTGAMMAAUTO and other options by default was used.
[3] The program FastME is published in: V. Lefort, R. Desper, and O. Gascuel. Molecular Biology and Evolution 32(10), 2015. The version 2.0.7 with default options was used. Input distances were calculated with the program *protdist* of the PHYLIP package with default options.

Funding: Russian Science Foundation, grant # 16-14-10319