

UFBoot2: Improving the Ultrafast Bootstrap Approximation

Diep Thi Hoang,^{†,1} Olga Chernomor,^{†,2} Arndt von Haeseler,^{2,3} Bui Quang Minh,^{*,2} and Le Sy Vinh^{*,1}

¹Faculty of Information Technology, University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam

²Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories, University of Vienna, Medical University Vienna, Vienna, Austria

³Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, Vienna, Austria

[†]These authors contributed equally to this work.

***Corresponding authors:** E-mails: minh.bui@univie.ac.at; vinhls@vnu.edu.vn.

Associate editor: Michael S. Rosenberg

Abstract

The standard bootstrap (SBS), despite being computationally intensive, is widely used in maximum likelihood phylogenetic analyses. We recently proposed the ultrafast bootstrap approximation (UFBoot) to reduce computing time while achieving more unbiased branch supports than SBS under mild model violations. UFBoot has been steadily adopted as an efficient alternative to SBS and other bootstrap approaches. Here, we present UFBoot2, which substantially accelerates UFBoot and reduces the risk of overestimating branch supports due to polytomies or severe model violations. Additionally, UFBoot2 provides suitable bootstrap resampling strategies for phylogenomic data. UFBoot2 is 778 times (median) faster than SBS and 8.4 times (median) faster than RAxML rapid bootstrap on tested data sets. UFBoot2 is implemented in the IQ-TREE software package version 1.6 and freely available at <http://www.iqtree.org>.

Key words: phylogenetic inference, ultrafast bootstrap, maximum likelihood, model violation, polytomies.

Standard nonparametric bootstrap (SBS) (Efron 1979; Felsenstein 1985) is widely used in maximum likelihood (ML) phylogenetic analyses to estimate branch supports of a phylogenetic tree inferred from a multiple sequence alignment (MSA). To achieve this, SBS generates a large number of resampled MSAs and reconstructs an ML-tree for each bootstrapped MSA. The resulting bootstrap ML trees are then used either to compute branch supports for the ML-tree reconstructed from the original MSA or to build a consensus tree with support values.

Although fast ML-tree search algorithms exist for large data sets (Vinh and von Haeseler 2004; Stamatakis 2006; Guindon et al. 2010; Nguyen et al. 2015) SBS is still very computationally intensive. To improve computing time, rapid bootstrap (RBS; Stamatakis et al. 2008) and the ultrafast bootstrap (UFBoot; Minh et al. 2013) were developed. Although RBS resembles the conservative behavior of SBS (i.e., underestimating branch supports), UFBoot provides relatively unbiased bootstrap estimates under mild model misspecifications.

The key idea behind UFBoot is to keep trees encountered during the ML-tree search for the original MSA and to use them to evaluate the tree likelihoods for the bootstrap MSAs. To speed up likelihood computation even further for bootstrap MSAs, IQ-TREE employed the resampling estimated log-likelihood (RELL) strategy (Kishino et al. 1990). For each bootstrap MSA, the tree with the highest RELL score (RELL-tree) represents the ML-bootstrap tree. Contrary to SBS, UFBoot does not further ML optimize this tree. The discrepancy in branch supports between UFBoot and SBS emerges as bootstrap trees inferred by UFBoot and SBS might be different.

Here, we present UFBoot2 that substantially speeds up UFBoot and reduces the risk for overestimated branch support due to polytomies or severe model violations. We also discuss several resampling strategies for phylogenomics data recently implemented in UFBoot2. In the following, we will outline these improvements.

Accelerating UFBoot

The likelihood computation is the major runtime bottleneck of all ML software because it lies at the core of all analyses. The pruning algorithm (Felsenstein 1981) efficiently computes the likelihood of phylogenetic trees, but still does not scale well for large data sets. Therefore, we adopted a modification to Felsenstein's algorithm (see supplementary method, Supplementary Material online), first introduced in RAxML. The modification exploits the reversible property of models of sequence evolution typically used in phylogenetic analysis, which led to a theoretical speedup of 4 (for DNA) or 20 (for protein data) when estimating branch lengths. Moreover, we employed the SIMD (single instruction, multiple data) feature to concurrently compute the likelihood of two MSA sites with streaming SIMD extensions or four MSA sites with advanced vector extensions, thus leading to a theoretical speedup of two or four compared with a non-SIMD implementation. IQ-TREE code was further optimized to avoid redundant computations.

We benchmark the runtimes on 70 DNA and 45 protein MSAs (DOI 10.5281/zenodo.854445) from TreeBASE, previously analyzed in Nguyen et al. (2015). The command-lines used to perform bootstrap methods are provided in

© The Author 2017. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Open Access

supplementary table S1, Supplementary Material online. UFBoot2 achieved a median speedup of 2.4 times (maximum: 77.3) compared with UFBoot version 0.9.6 (released on October 20, 2013).

Correction for Polytomies

Polytomies refer to multifurcating nodes in the tree that cannot be resolved due to low phylogenetic signal in the data. However, phylogenetic reconstruction always assumes strictly bifurcating trees. When resolving polytomies, there might be multiple equivalently optimal bifurcating trees (Whelan and Money 2010). As UFBoot (and other bootstrap approaches) saves only a single optimal bifurcating tree for each bootstrap MSA, it might cause overoptimistic bootstrap supports for short branches (Simmons and Norton 2014).

To correct for this shortcoming, UFBoot2 implemented the following technique. Instead of assigning the bootstrap tree with the highest RELL for each bootstrap MSA, UFBoot2 will randomly select one of the trees encountered during tree search, whose RELL score is less than ϵ_{boot} (default: 0.5) away from the highest RELL. As a result, UFBoot2 will not give high supports for branches resolving the multifurcations.

It was shown with a star tree simulation that SBS and RBS sometimes led to false positives (bootstrap supports $\geq 95\%$ for nonexisting branches), whereas with this technique UFBoot never supported such branches (support values $\leq 88\%$) (Simmons and Norton 2014). We repeated the star tree simulation for UFBoot2 with the same setting as proposed in (Simmons and Norton 2014). We used Seq-Gen 1.3.2x (Rambaut and Grass 1997) to evolve 100 DNA MSAs, each of 15,000 sites, along a 4-taxon star tree with four terminal branch lengths of 0.05, under JC model. For each MSA, we performed UFBoot2 runs under JC and GTR+ Γ , each with 1,000 bootstrap replicates and up to 1,000 search iterations (invoked in IQ-TREE via “-bcor 1” option). The simulation results show that UFBoot2 resembles the original UFBoot in that it never supports nonexisting branches (support values $\leq 88\%$).

Reducing the Impact of Model Violations

Minh et al. (2013) showed that severe model violations inflate UFBoot support values. To resolve this issue, UFBoot2 provides an option to conduct an additional step once the tree search on the original MSA is completed. Here, the best RELL-trees are further optimized using a hill-climbing nearest-neighbor interchange (NNI) search based directly on the corresponding bootstrap MSA. Thus, this extra step operates like SBS, but with a quick tree search to save time. Bootstrap supports are then summarized from the resulting corrected bootstrap trees. In the following, we called this UFBoot2 + NNI, which can be invoked in IQ-TREE via “-bnni” option.

We repeated the PANDIT simulations (Minh et al. 2013) to compare the accuracy of UFBoot2 and UFBoot2 + NNI with SBS (1,000 replicates using IQ-TREE) and RBS (RAXML bootstopping criterion). The simulations include 5,690 DNA MSAs (DOI 10.5281/zenodo.854445) generated by Seq-Gen

(Rambaut and Grass 1997), where the model parameters and the tree (which we will call the true tree in the following) were inferred from the original MSAs downloaded from the PANDIT database (Whelan et al. 2006). The accuracy of a bootstrap method M is defined by $f_M(x)$, the percentage of branches with support value x (across all reconstructed trees) that occur in the true tree (Hillis and Bull 1993). Thus, $f_M(x)$ reflects the probability that a branch with support x is a true branch. Figure 1 shows the results [y -axis depicts $f_M(x)$]. If the sequence evolution model used to infer the ML-tree agrees with the model used for simulations, then SBS, RBS, and UFBoot2 + NNI underestimated branch supports, the latter to a lower degree (fig. 1A; curves above the diagonal). This conservative behavior of SBS and RBS corroborates previous studies (Hillis and Bull 1993; Minh et al. 2013). Whereas UFBoot2 obtained almost unbiased branch supports (fig. 1A; curve close to the diagonal), that is, closely matching the true probability of branches being correct. Thus, UFBoot2 resembles the behavior of the original UFBoot (Minh et al. 2013).

Severe model violations do not influence SBS (fig. 1B; RBS not shown because RAXML does not support simpler models). However, UFBoot2 (like UFBoot) overestimated the branch supports (fig. 1B; curve below the diagonal), whereas UFBoot2 + NNI only slightly underestimated the bootstrap values (fig. 1B; curve closest to the diagonal). Thus, UFBoot2 + NNI helps to overcome the problem of unduly high supports by UFBoot2 in the presence of severe model violations.

In terms of computation times based on the analysis of 115 benchmark MSAs, UFBoot2, and UFBoot2 + NNI showed a median speedup of 778 (range: 200–1,848) and 424 (range: 233–749) compared with SBS, respectively. Compared with RBS, UFBoot2, and UFBoot2 + NNI are 8.4 (range: 1.5–51.2) and 5.0 (range: 0.8–32.6) times faster, respectively. Therefore, UFBoot2 + NNI is two times (median) slower than UFBoot2. Supplementary Figures S1–S3, Supplementary Material online, show the distributions of runtime ratios between SBS/RBS/UFBoot and UFBoot2/UFBoot2 + NNI.

We conclude that UFBoot2 and UFBoot2 + NNI are fast alternatives to other bootstrap approaches. Under no or mild model violations, UFBoot2 has the interpretation of unbiased bootstrap support as suggested for UFBoot (Minh et al. 2013). That is, one can trust branches with UFBoot2 support $\geq 95\%$. Users are advised to apply model violation detection methods (Goldman 1993; Weiss and von Haeseler 2003; Nguyen et al. 2011) before bootstrap analyses. UFBoot2 + NNI should be applied if severe model violations are present in the data set at hand.

Resampling Strategies for Phylogenomic Data

Recent phylogenetic analyses are typically based on multiple genes to infer the species tree, the so-called phylogenomics. To facilitate phylogenomic analysis, UFBoot2 implements several bootstrap resampling strategies: i) resampling MSA-sites within partitions (denoted as MSA-site resampling as the default option), ii) resampling genes instead of MSA-sites

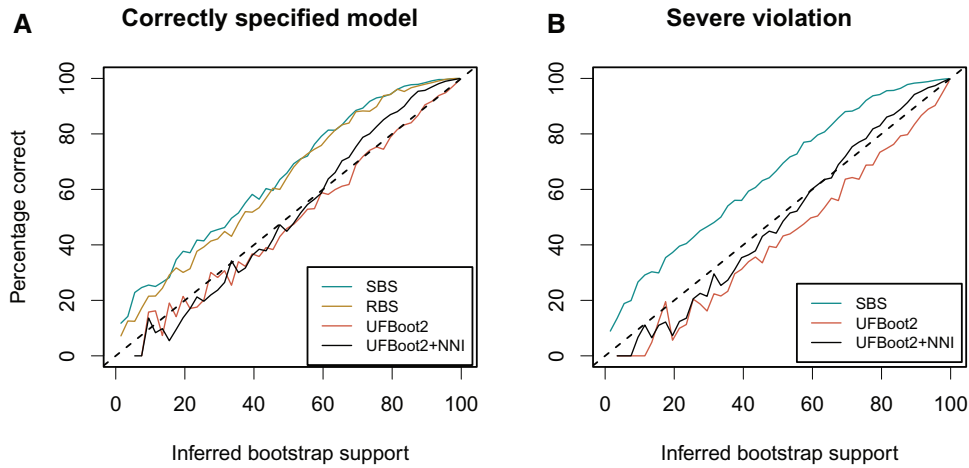


Fig. 1. Accuracy of the standard bootstrap (SBS), RAxML rapid bootstrap (RBS), ultrafast bootstrap (UFBoot2) and UFBoot2 with correction (UFBoot2 + NNI) for (A) correctly specified trees and (B) severely misspecified trees. The y-axis depicts the percentage of all branches with support value x (across all reconstructed trees) that occur in the true tree. Curves above the diagonal indicate underestimation of branch supports whereas curves below the diagonal indicate overestimation. For each point (x, y) in the curve representing the accuracy of bootstrap method M , x is an inferred bootstrap value by method M whereas y measures the probability of branches assigned by M with support value x to be true branches, that is, occurring on the true tree. Specifically, let X^+ (X^-) be the set of branches with support value x in all trees and present (absent) in the true tree. The y value is computed as the ratio between $|X^+|$ and $|X|$, where $|X| = |X^+| + |X^-|$.

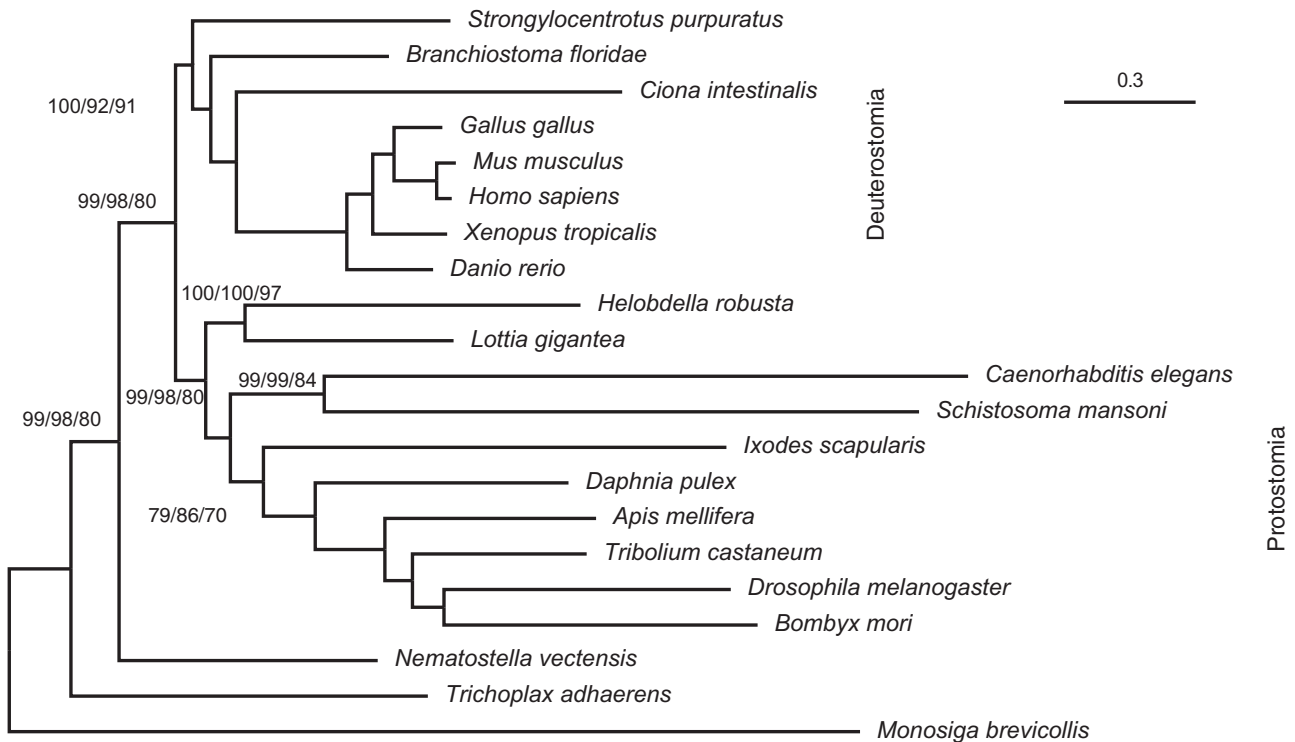


Fig. 2. Maximum-likelihood tree inferred under the edge-unlinked partition model. Numbers attached to the branches show the UFBoot2 bootstrap supports using MSA-site, gene, and gene-site resampling strategies (omitted when all three supports are 100%).

(gene-resampling invoked via “-bsam GENE” option), and iii) resampling genes and subsequently resamples MSA-sites within each gene (gene-site resampling, invoked via “-bsam GENESITE” option) (Gadagkar et al. 2005). Strategy (i) preserves the number of MSA-sites for all genes in the bootstrap MSAs, whereas strategies (ii) and (iii) will lead to different number of sites in the bootstrap MSAs.

To investigate the impact of the three resampling strategies, we reanalyzed the metazoan data with 21 species, 225 genes, and a total of 171,077 amino-acid sites (Salichos and Rokas 2013). Figure 2 shows the ML tree inferred with IQ-TREE under edge-unlinked partition model (Chernomor et al. 2016), which allows separate sets of branch lengths across partitions. The tree replicates previous results (Salichos and

Rokas 2013) and shows the Protostomia clade (Telford et al. 2015). However, discrepancies between resampling strategies are observed: while MSA-site and gene-resamplings obtained high supports (>95%) for branches along the backbone of the tree (fig. 2; bold lines), lower supports (80%) were estimated by gene-site resampling.

By further examining 14 other empirical data sets (Bouchenak-Khelladi et al. 2008; Fabre et al. 2009; Stamatakis and Alachiotis 2010; van der Linde et al. 2010; Pyron et al. 2011; Nyakatura and Bininda-Emonds 2012; Springer et al. 2012; Hinchliff and Roalson 2013; Salichos and Rokas 2013; Dell’Ampio et al. 2014), we observed more discrepancies between resampling strategies (data not shown). Exceptionally, for some data sets, a number of branches showed almost no support ($\leq 10\%$) for one resampling but high supports ($\geq 95\%$) for the other two resampling strategies. However, there is no tendency toward systematically lower supports obtained by one resampling strategy.

Taking into account the above findings, we recommend to apply all alternative resampling strategies. If similar bootstrap supports are obtained, then one can be more confident about the results.

Conclusions

UFBoot2 significantly improves speed and accuracy of bootstrap values compared with UFBoot. It also offers new functionalities in the presence of model violations and in its applicability to phylogenomic data. In general, since SBS, RBS, and UFBoot2 + NNI share a disadvantage of being conservative, more research is necessary to understand the different biases introduced by the available phylogenetic bootstrap estimation methods.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

This work was supported by Vietnam National Foundation for Science and Technology Development (102.01-2013.04). A.V.H., B.Q.M., and O.C. were supported by the Austrian Science Fund—FWF (grant nos. I-2805-B29 and I-1824-B22). The authors thank Stephen Crotty and two anonymous reviewers for helpful comments on the manuscript.

References

- Bouchenak-Khelladi Y, Salamin N, Savolainen V, Forest F, van der Bank M, Chase MW, Hodkinson TR. 2008. Large multi-gene phylogenetic trees of the grasses (Poaceae): progress towards complete tribal and generic level sampling. *Mol Phylogenet Evol.* 47(2):488–505.
- Chernomor O, von Haeseler A, Minh BQ. 2016. Terrace aware data structure for phylogenomic inference from supermatrices. *Syst Biol.* 65(6):997–1008.
- Dell’Ampio E, Meusemann K, Szucsich NU, Peters RS, Meyer B, Borner J, Petersen M, Aberer AJ, Stamatakis A, Walz MG, et al. 2014. Decisive data sets in phylogenomics: lessons from studies on the phylogenetic relationships of primarily wingless insects. *Mol Biol Evol.* 31(1):239–249.
- Efron B. 1979. Bootstrap methods: another look at the jackknife. *Ann Stat.* 7(1):1–26.
- Fabre P-H, Rodrigues A, Douzery EJP. 2009. Patterns of macroevolution among Primates inferred from a supermatrix of mitochondrial and nuclear DNA. *Mol Phylogenet Evol.* 53(3):808–825.
- Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol.* 17(6):368–376.
- Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution (NY)* 39(4):783–791.
- Gadagkar SR, Rosenberg MS, Kumar S. 2005. Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. *J Exp Zool B Mol Dev Evol.* 304B(1):64–74.
- Goldman N. 1993. Statistical tests of models of DNA substitution. *J Mol Evol.* 36(2):182–198.
- Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol.* 59(3):307–321.
- Hillis DM, Bull JJ. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst Biol.* 42(2):182–192.
- Hinchliff CE, Roalson EH. 2013. Using supermatrices for phylogenetic inquiry: an example using the sedges. *Syst Biol.* 62(2):205–219.
- Kishino H, Miyata T, Hasegawa M. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J Mol Evol.* 31(2):151–160.
- Minh BQ, Nguyen MAT, von Haeseler A. 2013. Ultrafast approximation for phylogenetic bootstrap. *Mol Biol Evol.* 30(5):1188–1195.
- Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol.* 32(1):268–274.
- Nguyen MAT, Klaere S, von Haeseler A. 2011. MISFITS: evaluating the goodness of fit between a phylogenetic model and an alignment. *Mol Biol Evol.* 28(1):143–152.
- Nyakatura K, Bininda-Emonds ORP. 2012. Updating the evolutionary history of Carnivora (Mammalia): a new species-level super-tree complete with divergence time estimates. *BMC Biol.* 10(1):12.
- Pyron RA, Burbrink FT, Colli GR, de Oca ANM, Vitt LJ, Kuczynski CA, Wiens JJ. 2011. The phylogeny of advanced snakes (Colubroidea), with discovery of a new subfamily and comparison of support methods for likelihood trees. *Mol Phylogenet Evol.* 58(2):329–342.
- Rambaut A, Grass NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics* 13(3):235–238.
- Salichos L, Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497(7449):327–331.
- Simmons MP, Norton AP. 2014. Divergent maximum-likelihood-branch-support values for polytomies. *Mol Phylogenet Evol.* 73:87–96.
- Springer MS, Meredith RW, Gatesy J, Emerling CA, Park J, Rabosky DL, Stadler T, Steiner C, Ryder OA, Janečka JE, et al. 2012. Macroevolutionary dynamics and historical biogeography of primate diversification inferred from a species supermatrix. Stanyon R, editor. *PLoS One* 7:e49521.
- Stamatakis A. 2006. RAXML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22(21):2688–2690.
- Stamatakis A, Alachiotis N. 2010. Time and memory efficient likelihood-based tree searches on phylogenomic alignments with missing data. *Bioinformatics* 26(12):i132–i139.
- Stamatakis A, Hoover P, Rougemont J, Renner S. 2008. A rapid bootstrap algorithm for the RAXML web servers. *Syst Biol.* 57(5):758–771.
- Telford MJ, Budd GE, Philippe HH. 2015. Phylogenomic insights into animal evolution. *Curr Biol.* 25(19):R876–R887.

- van der Linde K, Houle D, Spicer GS, Stepan SJ. 2010. A supermatrix-based molecular phylogeny of the family Drosophilidae. *Genet Res (Camb)* 92(1):25–38.
- Vinh LS, von Haeseler A. 2004. IQPNNI: moving fast through tree space and stopping in time. *Mol Biol Evol.* 21(8):1565–1571.
- Weiss G, von Haeseler A. 2003. Testing substitution models within a phylogenetic tree. *Mol Biol Evol.* 20(4):572–578.
- Whelan S, de Bakker PIW, Quevillon E, Rodriguez N, Goldman N. 2006. PANDIT: an evolution-centric database of protein and associated nucleotide domains with inferred trees. *Nucleic Acids Res.* 34:D327–D331.
- Whelan S, Money D. 2010. The prevalence of multifurcations in tree-space and their implications for tree-search. *Mol Biol Evol.* 27(12):2674–2677.