How Many Bootstrap Replicates Are Necessary?

Nicholas D. Pattengale[1], Masoud Alipour[2], Olaf R.P. Bininda-Emonds[3], Bernard M.E. Moret[2,4], and Alexandros Stamatakis[5]

[1] Department of Computer Science, University of New Mexico, Albuquerque NM, USA
[2] Laboratory for Computational Biology and Bioinformatics, EPFL (École Polytechnique Fédérale de Lausanne), Switzerland
[3] AG Systematik und Evolutionsbiologie, Institut für Biologie und Umweltwissenschaften, University of Oldenburg, Germany
[4] Swiss Institute of Bioinformatics, Lausanne, Switzerland
[5] The Exelixis Lab, Department of Computer Science, Technische Universität München, Germany

nickp@cs.unm.edu, masoud.alipour@epfl.ch, olaf.bininda@uni-oldenburg.de, bernard.moret@epfl.ch, stamatak@cs.tum.edu

http://icwww.epfl.ch/~stamatak/index-Dateien/Page443.htm

Abstract. Phylogenetic Bootstrapping (BS) is a standard technique for inferring confidence values on phylogenetic trees that is based on reconstructing many trees from minor variations of the input data, trees called replicates. BS is used with all phylogenetic reconstruction approaches, but we focus here on the most popular, Maximum Likelihood (ML). Because ML inference is so computationally demanding, it has proved too expensive to date to assess the impact of the number of replicates used in BS on the quality of the support values. For the same reason, a rather small number (typically 100) of BS replicates are computed in real-world studies. Stamatakis *et al.* recently introduced a BS algorithm that is 1–2 orders of magnitude faster than previous techniques, while yielding qualitatively comparable support values, making an experimental study possible.

In this paper, we propose *stopping criteria*, that is, thresholds computed at runtime to determine when enough replicates have been generated, and report on the first large-scale experimental study to assess the effect of the number of replicates on the quality of support values, including the performance of our proposed criteria. We run our tests on 17 diverse real-world DNA, single-gene as well as multi-gene, datasets, that include between 125 and 2,554 sequences. We find that our stopping criteria typically stop computations after 100–500 replicates (although the most conservative criterion may continue for several thousand replicates) while producing support values that correlate at better than 99.5% with the reference values on the best ML trees. Significantly, we also find that the stopping criteria can recommend very different numbers of replicates for different datasets of comparable sizes.

Our results are thus two-fold: (i) they give the first experimental assessment of the effect of the number of BS replicates on the quality of support values returned through bootstrapping; and (ii) they validate

S. Batzoglou (Ed.): RECOMB 2009, LNCS 5541, pp. 184–200, 2009.

our proposals for stopping criteria. Practitioners will no longer have to enter a guess nor worry about the quality of support values; moreover, with most counts of replicates in the 100–500 range, robust BS under ML inference becomes computationally practical for most datasets. The complete test suite is available at `http://lcbb.epfl.ch/BS.tar.bz2` and BS with our stopping criteria is included in RAxML 7.1.0.

Keywords: Phylogenetic Inference, Maximum Likelihood, Bootstrap, Support Value, Stopping Criterion, Bootstopping.

1 Introduction

Phylogenetic trees are used to represent the evolutionary histories of related organisms (as well, of course, as of any other units subject to evolutionary changes, from protein through genes and genomes to languages and ecologies). Most phylogenetic reconstructions for a collection of organisms take as input DNA or protein sequence alignments. (Others may take encoded morphological characters, although the end result remains a collection of aligned sequences.) These input sequences are placed at the leaves of the putative tree and reconstruction proceeds by searching for an optimal internal branching structure for the tree. Due to the rapid, and rapidly accelerating, growth of sequence data in the last few years, reconstruction of trees with more than 1,000 leaves has become increasingly common, often using sequence data from many genes (so-called multi-gene or phylogenomic alignments). Such practice represents a major departure from the typical practice of the last 20 years, in which trees of 10-100 organisms were inferred from the sequence of a few simple ribosomal genes. Scaling up inference in terms of the number of organisms, the length and complexity of the sequence data, and the diameter (largest pairwise distance among the organisms) is a very challenging issue [19]. The search space (all possible distinct branching structures) is notoriously large ($(2n-5)!! = (2n-5)\cdot(2n-7)\ldots 5\cdot 3\cdot 1$) and unstructured. Both Maximum Parsimony and Maximum Likelihood approaches are known to be NP-hard, but both are preferred to the simpler distance methods, especially in the presence of more complex data or data with large diameters.

Significant progress has been achieved in the field of heuristic ML search algorithms with programs such as PHYML [12], GARLI [33], LeaPhy [32], and RAxML [29]. However, there is still a major bottleneck in computing bootstrap support (BS) values on these trees, which can require more than one month of sequential execution time for a likely insufficient number of 100 replicates [28] on a reasonably fast CPU. To date, it has proved infeasible to assess empirically the convergence properties of BS values, much less to evaluate means for dynamically deciding when a set of replicates is sufficiently large—at least on the size of trees where computing BS values is an issue.

Recently, Stamatakis *et al.* [30] introduced a fast BS algorithm that yields a run time acceleration of one to two orders of magnitude compared to other current algorithms while returning qualitatively comparable support values. This

improvement makes possible a large-scale experimental study on bootstrap stopping criteria, the results of which are the topic of this paper.

We propose two stopping criteria. Both split the set of replicates computed so far into two equal sets and compute statistics on the two sets. The frequency criterion (FC) is based on the observed frequencies of occurrences of distinct bipartitions; the more conservative weight criterion (WC) computes the consensus tree for each subset and scores their similarity. Both criteria can be computed efficiently and so a stopping test can be run every so many replicates until stopping is indicated. We test these criteria and the general convergence properties of BS values on 17 diverse real-world DNA, single-gene, as well as multi-gene datasets, that include between 125 and 2,554 sequences. We find that our stopping criteria typically stop computations after 100–500 replicates (although the most conservative criterion may continue for several thousand replicates) while producing support values that correlate at better than 99.5% with the reference values on the best ML trees. Unsurprisingly, differences tend to occur mostly on branches with poor support—on branches with support values of at least 0.75, over 98% of the values returned after early stopping agree with the reference values to within 5%.

Our results show that the BS convergence speeds of empirical datasets are highly dataset-dependent, which means that bootstopping criteria can and should be deployed to determine convergence on a per alignment basis. The criteria help to conduct as many BS replicates as *necessary* for a given accuracy level and thus help to reduce the computational costs for phylogenetic analyses. Practitioners will no longer have to enter a guess nor worry about the quality of support values; moreover, with most counts of replicates in the 100–500 range, robust BS under ML inference becomes computationally practical for most datasets.

The remainder of this paper is organized as follows: In Section 2, we review the bootstrap concept and related work on stopping criteria for (mostly non-phylogenetic) bootstrap procedures, including a brief overview of convergence criteria for MrBayes [26]. In Section 3 we describe our family of stopping criteria. In Section 4 we describe our experimental study, give detailed results, and discuss their implications.

2 Related Work on Bootstopping Criteria

2.1 The Phylogenetic Bootstrap

Phylogenetic bootstrapping is a fairly straightforward application of the standard statistical (nonparametric) bootstrap and was originally suggested by Felsenstein [9] as a way to assign confidence values to edges/clades in phylogenetic trees. Phylogenetic BS proceeds by generating perturbed BS alignments which are assembled by randomly drawing alignment columns from the original input alignment with replacement. The number of columns in the bootstrapped alignment is identical to the number of columns in the original alignment, but the column composition is different. Then, for each BS alignment, a tree is reconstructed independently. The procedure returns a collection of tree *replicates*.

The replicates can then be used either to compute consensus trees of various flavors or to draw confidence values onto a reference tree, e.g., the best-scoring ML tree. Each edge in such a reference tree is then assigned a confidence value equal to the number of replicates in which it appears. The question we address in this paper is—how many replicates must be generated in order to yield accurate confidence values? By accurate confidence values we mean relative accuracy of support values (the "true" support values are unknown for empirical datasets) with respect to support values obtained by a very large number ($\geq$ 10,000 in our experiments) of reference replicates. The extent to which the question about the appropriate number of BS replicates has been answered in other applications of the (non-phylogenetic) bootstrap is the subject of the following subsection.

2.2 General Bootstopping Criteria

Most of the literature addressing (whether theoretically or empirically) the issue of ensuring a sufficient number of replicates stems from the area of general statistics or econometrics. However, they are difficult to apply to phylogenetic BS due to the significantly higher computational and theoretical complexity of the estimator [17]. In addition, the problem is more complex since the number of entities (bipartitions) to which support values are assigned grows during the BS procedure, i.e., adding more BS replicates increases the number of unique bipartitions. This is not commonly the case for other application areas of the general Bootstrapping procedure and general bootstopping criteria that have recently been proposed (for instance see [13]).

Standard textbooks on Bootstrapping such as [6,8] suggest to choose a sufficiently large number B of BS replicates without addressing exact bounds for B. This does not represent a problem in most cases where the BS procedure is applied to simple statistical measures such as the mean or variance of univariate statistics. Efron and Tibshirani [8] suggest that $B = 500$ is sufficient for the general standard bootstrap method in most cases. Manly *et al* [18] propose a simple approach to determine B *a priori*, i.e., before conducting the BS analysis, based on a worst-case scenario by approximating the standard deviation of BS statistics. The analysis in [18] concludes that a general setting of $B = 200$ provides a relatively small error margin in BS estimation. This approximation can only be applied to standard BS procedures, based on simple, univariate statistics. However, a larger number of BS replicates is required for other applications of the Bootstrap such as the computation of confidence intervals or tests of significance. P. Hall [14] proposes a general method for stopping the BS in a percentile-t confidence interval. In the area of econometrics, Davidson and MacKinnon [7] propose a two-step procedure to determine B for BS P-values based on the most powerful test. Andrews *et al.* [1,2,3] propose and evaluate a general three-step algorithm to specify B in the bootstrap procedure. Andrews and Buchinsky [4] then further extend their algorithm to bootstrap BCA intervals.

With respect to phylogenetics Hedges [15] suggests a method to specify B *a priori* for a given level of significance. This approach does not take into account the number of sequences and hence the number of potential alternative

tree topologies, or the number of base-pairs or distinct patterns in the alignment. However, as underlined by our experimental results, important alignment-specific properties such as the "gappyness" (percentage of gaps) of the alignment, the quality of the alignment, and the respective phylogenetic signal strength greatly influence the estimator (the tree search algorithm) and hence the stability of BS replicates. We conclude that an adaptive stopping criterion which is computed on the fly at regular intervals during the actual BS search is best suited to take into account the particularities of real-world datasets and to determine a useful trade-off between accuracy and inference time. We are convinced that such trade-offs will become increasingly important for analysis on large phylogenomic datasets under computational resource constraints, as a current collaborative study with Biologists already requires 2,000,000 CPU hours on an IBM BlueGene/L supercomputer. Therefore, we assess our approach empirically, via a large number of computational experiments on diverse real datasets.

2.3 Bayesian Convergence Criteria and Tools

There exists some work on convergence criteria and tools for Bayesian phylogenetic analyses, most probably because the convergence of the actual search as opposed to a sufficient number of BS replicates in ML represents a more serious methodological problem for MCMC in general and phylogenetic MCMC searches in particular [20,27,31]. Gelman, Rubin, and Brooks [5,10] provide general frameworks to determine convergence of iterative simulations, with a focus on MCMC methods. MrBayes implements convergence diagnostics for multiple Metropolis-coupled MCMC chains that use the average standard deviation in partition frequency values across independent analyses. One potential drawback is that these statistics take into account all partition frequencies and not only the important, highly supported ones. In addition, there exist tools for graphical exploration of convergence such as AWTY [21] to visualize convergence rates of posterior split probabilities and branch lengths or Tracer [23] that analyzes time-series plots of substitution model parameters. AWTY also offers bivariate plots of split frequencies for trees obtained via independent chains. Note that both AWTY and Tracer require the user to visually inspect the respective output and determine whether the MCMC chains have converged. We are not aware of any computational experiments to assess the performance and accuracy of the above methods.

3 Bootstopping Criteria

In this section, we introduce stopping criteria for bootstrapping procedures, which we call "bootstopping" criteria. These are measures that are computed and used at run time, during the replicate generation phase, to decide when enough replicates have been computed. The frequency-based criterion (FC) is based upon Pearson's correlation coefficient, whereas the Weighted Robinson-Foulds criterion (WC) is based upon the (weighted) symmetric topological difference widely used in phylogenetics.

3.1 Terminology and Definitions

A phylogenetic tree T is an unrooted binary tree; its leaves (also called tips) are labelled by the organism names of the input alignment, while its internal nodes represent hypothetical extinct common ancestors. Removing a branch between nodes a and b from a tree T disconnects the tree and creates two smaller trees, T_a and T_b. The trees T_a and T_b induce a *bipartition* (or split) of the set S of taxa (organism names at the leaves) of T into two disjoint taxon sets A and B ($A \cup B = S$). We denote such a bipartition as $A|B$. Thus, there exists a one-to-one correspondence between the bipartitions of S and the branches of T, so that each tree is uniquely characterized by the set of bipartitions it induces. If $|S| = n$, then any (unrooted) multifurcating phylogenetic tree for S has at most $2n - 3$ branches and so induces at most $2n - 3$ bipartitions. If the tree is fully bifurcating the number of bipartitions is exactly $2n - 3$, while the number of non-trivial bipartitions, i.e., splits at branches that do not lead to a tip, is $n - 3$.

The Robinson-Foulds (RF) metric (sometimes referred to as symmetric difference) is a dissimilarity metric between two trees and counts the number of bipartitions that occur in one tree and not the other. The Weighted Robinson-Foulds (WRF) metric generalizes the RF metric by summing the weights of the bipartitions that contribute to the RF metric (and also, optionally, includes the sum of differences between the weights of shared bipartitions). Finally, consensus methods take a set of trees and return a single 'summary' tree. The majority rule consensus method (MR) returns a tree containing only bipartitions that exist in greater than half the input trees. The extended majority rules method (MRE, also known as *greedy consensus*) uses the MR consensus tree as a starting point and greedily adds bipartitions that occur in less than half the input trees by descending order of their frequency in the hopes (although not always possible) of obtaining a fully bifurcating (binary) tree.

3.2 Stopping Criteria

The two criteria we present in the following are both based on the same underlying mechanism. Initially, the set of replicates to be tested for convergence is randomly split into two equal halves. Then we compute statistics between the bipartition support values induced by these halves. If the difference between the splits of the replicates are small this indicates that adding more replicates will not significantly change the bipartition composition of the replicate set. In addition, we compute the statistics not only for one but for 100 random splits of the replicate sets, i.e., we draw a sample from all possible random splits of the replicates by applying a permutation test.

Frequency Criterion (FC) The frequency-based criterion uses the bipartition frequencies of all replicates computed up to the point at which the test is conducted, for example every 50 replicates, i.e., at 50, 100, 150, 200, ... replicates. One major design goal is to devise stand-alone criteria that do not rely on a previously computed best-known ML tree for the original alignment. This is partially due to the rapid BS algorithm (and future extensions thereof) in

RAxML that uses information gathered during the BS search to steer and accelerate the search for the best-scoring ML tree on the original alignment. Another important goal is to avoid a heavy dependency on the spacing (e.g., every 10, 20, or 50 replicates) of two successive steps of the test, i.e., we do not want to compute statistics that compare 20 with 30 replicates. Therefore, we have adopted a procedure, that is in some sense similar to the aforementioned convergence tests for MCMC chains implemented in MrBayes. There are two main differences though: (i) we do not use the test to determine convergence of the tree search itself, and (ii) we do not apply the test to only one single random or fixed split of the replicate tree set.

Our FC test works as follows: Assume that the test is conducted every 50 replicates, i.e., after the computation of 50, 100, 150, ... BS replicates. This spacing of 50 has been chosen empirically, in order to achieve a reasonable computational trade-off between the cost of the test and the cost for computing replicates (future work will cover the development of adaptive spacing strategies). The empirical setting also fits the typical range of bootstopped tree topologies, which range between 150 and 450 in our FC-based experiments, depending on the strength of the signal in the respective alignment. For the sake of simplicity, assume that we conduct the test for 50 replicates. At the top level of our procedure we perform a permutation test by randomly splitting up those 50 tress p =100 times (p =100 permutations) into disjoint sets s_1,s_2 of equal size with 25 trees each. The advantage of 100 random splits over a single random split or a fixed split into, e.g., replicates with even and odd numbers, is that the curve is smoothed and depends to a far lesser degree on a by chance favorable or unfavorable single split of the data.

In Figure 1 we depict the impact of using p =1, 10, and 100 permutations on the FC and WC criteria (see Sect. 3.2) for a dataset with 500 sequences. As expected the curve becomes smoother for larger p settings; a setting of $p = 10$ appears to be sufficient to smooth the curve and reduce the cost of the test. Though statistically more stable, the disadvantage of this approach is clearly the significantly increased computational cost of the test. Nonetheless, an initial, highly optimized at a technical level, yet algorithmically naïve implementation requires only 1 minute to conduct all 6 tests on 50, 100, ..., 300 replicates on a 1,481 taxon dataset (2 minutes, 40 seconds for $p = 1000$ random splits), compared to roughly 27 hours for the computation of 300 rapid BS replicates.

For each of the aforementioned 100 random splits we compute the support vectors v_1 for s_1 and v_2 for s_2 for all bipartitions b_{ALL} found in $s_1 \cup s_2$, i.e., all bipartitions contained in the original 50 trees. Note that both vectors v_1, v_2 have length b_{ALL}. Given those two vectors for each permutation (random split) i, where, $i = 0, ..., 99$ we simply compute Pearson's correlation coefficient ρ_i on the vectors. Our procedure stops if there are at least 99 ρ_i with $\rho_i \geq 0.99$ (only one possible parameter setting). We henceforth denote the Pearson's threshold used as ρ_{FC}. A potential drawback of this method is that the support frequencies on the best-scoring tree or for all bipartitions found during the BS search might not follow a normal distribution. Nonetheless, the FC method appears to work

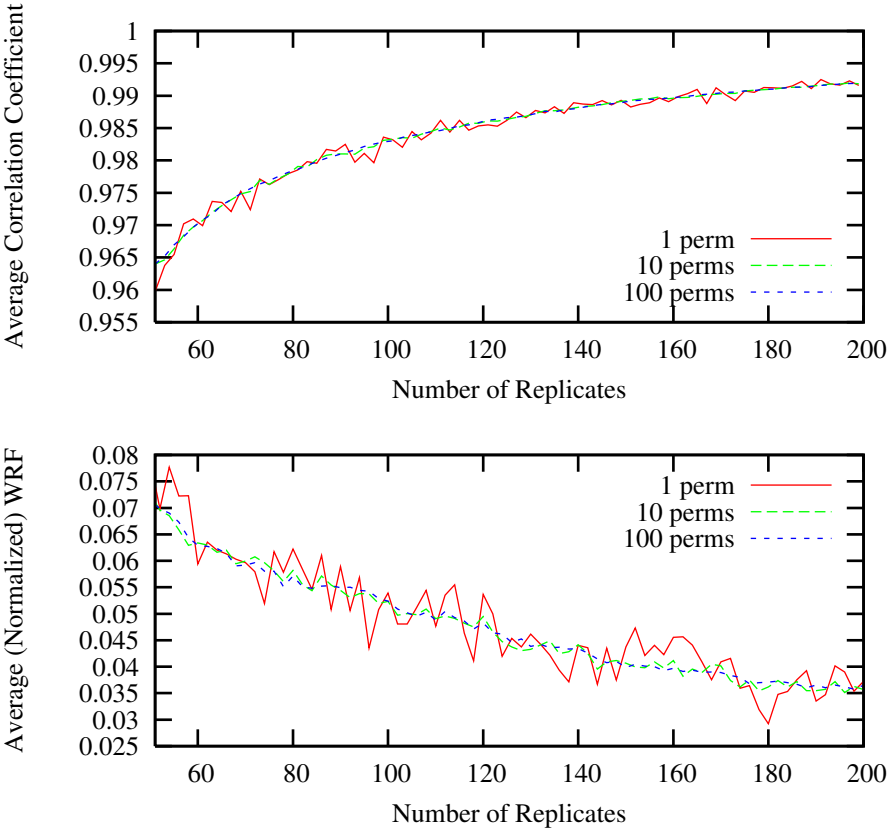

Fig. 1. FC (top) and WC (bottom) criteria for various p settings on dataset 500

reasonably well in practice (see Section 4). Another potential drawback is that the FC criterion is based on the bipartition frequencies of all bipartitions found. However, from a biological point of view, one is only interested in the "important" bipartitions, i.e., the bipartitions induced by the best-scoring ML tree or the bipartitions that form part of a strict, majority rule, or extended majority rule consensus tree. We address the design of a criterion that only takes into account important bipartitions in the next section. Nonetheless, the FC test can easily be extended in the future to take into account the important bipartitions by providing a user-defined best-scoring ML tree using either Pearson's correlation or, e.g., the mean square error between corresponding bipartition support values.

Weighted Robinson-Foulds distance-based Criterion (WC) The Weighted Robinson-Foulds (WRF) distance criterion (WC) is employed similarly to the FC criterion (i.e. every 50 trees and uses $p = 100$ permutations per test). Rather than computing a vector correlation, we compute the majority rules consensus trees for

s_1 and s_2 and then assess the (dis)similarity between the two consensus trees. We then use the respective consensus trees, which only contain support values for "important" biologically relevant partitions, to calculate the WRF distance between the consensus tree $c(s_1)$ of tree set s_1 and the consensus tree $c(s_2)$ of tree set s_2.

As a distance measure and hence convergence criterion we use the weighted Robinson-Foulds distance (WRF). This weighted topological distance measure between consensus trees takes into account the support values and penalizes incongruent subtrees with low support to a lesser extent. When RF distances are significantly larger than their weighted counterparts (WRF), this indicates that the differences in the consensus trees are induced by subtrees with low support. When WRF≈RF this means that the differences in the tree topologies under comparison are due to differently placed clades/subtrees with high support. From a biological perspective the WRF distance represents a more reasonable measure since systematists are typically interested in the phylogenetic position of subtrees with high support. In real-world studies the typical empirical threshold is set to 75%, i.e., clades with a BS support of $\geq 75\%$ are usually considered to be monophyletic (see [28] for a summary). As for the FC criterion, the WC stopping rule can be invoked with varying numbers of permutations and threshold settings. One might for example stop the BS procedure, if for $p = 99$ out of 100 permutations, the relative WRF between $c(s_1)$ and $c(s_2)$ is $\leq 5\%$. For reasons of consistency we also denote the threshold parameter for WC as ρ_{WC}, a ρ_{WC} setting of 0.97 means that the BS search is stopped when p WRF distances are $\leq 1.0 - 0.97 = 3\%$.

4 Experimental Setup and Results

4.1 Experimental Setup

To test the performance and accuracy of FC and WC we used 17 real-world DNA alignments containing 125 up to 2,554 sequences. The number of distinct alignment patterns ranges between 348 and 19,436. For the sake of simplicity, alignments will henceforth be referenced by the number of taxa as provided in Table 1. The experimental data spans a broad range of mostly hand-aligned sequences including rbcL genes (500, 2,554), mammalian sequences (125, 1,288, 2,308), bacterial and archaeal sequences (714, 994, 1,481, 1,512, 1,604, 2,000), ITS sequences (354), fungal sequences (628, 1,908), and grasses (404). The 10,000 reference BS replicates on each dataset were inferred on two AMD-based Linux clusters with 128 and 144 CPUs, respectively. All result files and datasets used are available for download at `http://lcbb.epfl.ch/BS.tar.bz2` We make this data available in the hope that it will be useful as a basis for further exploration of stopping criteria as well as general properties of BS.

Computational experiments were conducted as follows. For each dataset we computed a minimum of 10,000 BS replicates using the rapid Bootstrapping (RBS [30]) algorithm implemented in RAxML. We then applied stand-alone bootstopping tests (either FC or WC) that take the set of 10,000 BS reference replicates as input and only execute the tests described in Section 3 without

Table 1. Performance analysis of FC ($p = 99$, $\rho_{FC} = 0.99$) vs. WC ($p = 99$, $\rho_{WC} = 0.97$) for three metrics: number of trees to converge, WRF between MRE consensus trees and Correlation Coefficient. Column # Patterns indicates the number of distinct column patterns in each alignment. The last line depicts the respective averages.

DATA	CON-FC	CON-WC	WRF-FC	WRF-WC	P-FC	P-WC	# Patterns
125	150	50	0	0	0.9997	0.9994	19,436
150	250	650	0.03	0.01	0.9984	0.9994	1,130
218	300	550	0.04	0.01	0.9977	0.9988	1,846
354	450	1200	0.03	0.01	0.9979	0.9992	348
404	250	700	0.04	0.01	0.9965	0.9988	7,429
500	200	400	0.03	0.01	0.9982	0.9991	1,193
628	250	450	0.03	0.01	0.9975	0.9987	1,033
714	200	400	0.03	0.02	0.9977	0.9989	1,231
994	150	300	0.04	0.02	0.9964	0.9974	3,363
1,288	200	400	0.03	0.02	0.9967	0.9985	1,132
1,481	300	450	0.04	0.02	0.9968	0.9979	1,241
1,512	250	350	0.03	0.02	0.9977	0.9983	1,576
1,604	250	600	0.04	0.02	0.9975	0.9990	1,275
1,908	200	400	0.03	0.02	0.9975	0.9987	1,209
2,000	300	600	0.03	0.01	0.9976	0.9989	1,251
2,308	150	200	0.03	0.02	0.9980	0.9985	1,184
2,554	200	500	0.03	0.01	0.9975	0.9991	1,232
1,102	238	482	0.03	0.01	0.9976	0.9987	2,771

performing the actual BS search. Returned is a file containing the first k trees from the full set, where k is determined by the stopping criterion (FC or WC, along with appropriate parameter values). We refer to these first k trees as the 'bootstopped' trees.

We then computed a number of (dis)similarity metrics between the reference replicates and the bootstopped replicates, including: correlation coefficient, RF between MRE consensus trees of the two sets, and WRF between the MRE consensus trees of the two sets. Additionally, support values from the bootstopped and full replicate sets were drawn on the best-scoring ML tree and the resulting support values compared.

4.2 Results for FC and WC Methods

In Table 1 we provide basic performance data for FC and WC. Column *DATA* lists the alignments, *CON-FC* the FC bootstop convergence number, and column *CON-WC* the WC bootstop convergence number. Columns *WRF-FC* and *WRF-WC* provide the WRF distance between the MRE consensus tree for the bootstopped trees and the MRE consensus tree induced by the reference replicates for FC and WC respectively. Finally, columns *P-FC* and *P-WC* provide Pearson's correlation coefficient between support values from the bootstopped trees and the reference trees on the best-scoring ML tree for FC and WC respectively.

We observe that WC tends to be more conservative, i.e., stops the BS search after more replicates, except for dataset 125. Dataset 125 is a particularly long

Table 2. Performance analysis of FC ($p = 99$, $\rho_{FC} = 0.99$) vs. WC ($p = 99$, $\rho_{WC} = 0.97$) for three metrics: mean error, mean squared error, and loss of support. The last line depicts the respective averages.

DATA	μ-FC	σ^2-FC	μ-WC	σ^2-WC	SUPPLOSS-FC	SUPPLOSS-WC
125	0.303279	0.637530	0.483607	1.807108	0.001066	0.004672
150	1.544218	2.941922	1.074830	1.402564	0.009252	0.003605
218	1.865116	3.205062	1.297674	1.836971	0.005070	0.004674
354	1.364672	1.912598	0.886040	0.864506	0.002009	0.002835
404	2.553616	6.626178	1.384040	2.386179	0.012357	0.007170
500	1.792757	3.532503	1.239437	1.936634	0.010020	0.006841
628	2.030400	4.531876	1.398400	2.175677	0.013400	0.008408
714	2.129395	4.973412	1.424754	2.396237	0.010858	0.008833
994	2.498486	11.178353	2.068618	9.014464	0.013895	0.010575
1,288	2.477821	8.308652	1.700389	3.752257	0.013899	0.009864
1,481	1.845061	5.082219	1.496617	3.243223	0.008562	0.007287
1,512	1.762094	3.958643	1.552684	3.176317	0.008403	0.006289
1,604	1.898813	3.891073	1.229232	1.746953	0.008120	0.005721
1,908	1.961680	4.209030	1.377528	2.298479	0.009711	0.007113
2,000	1.773160	3.323105	1.184276	1.504350	0.008488	0.005020
2,308	1.951410	6.626706	1.703254	4.919317	0.010330	0.009681
2,554	2.063897	4.639194	1.248530	1.793192	0.011319	0.006370
1,102	1.871522	4.681062	1.338230	2.720849	0.009221	0.006762

phylogenomic alignment of mammals and exhibits a surprisingly low variability for the bipartitions it induces. The 10,000 reference replicates only induce a total of 195 distinct bipartitions, which is extremely low given that a single BS tree for this dataset induces $125 - 3 = 122$ nontrivial bipartitions. The WC method appears to capture this inherent stability of the BS trees sooner than FC, while the WRF to the MRE tree is 0 in both cases, i.e., the consensus trees for 50, 150, and 10,000 replicates are exactly identical. This also underlines our claim that our criteria help avoid needless computation (and needless energy expenditures, as large clusters tend to be power-hungry), in particular on such large and challenging phylogenomic datasets. Due to the general trend for WC to stop later, both *WC* metrics (P/WRF) are higher than the respective values for FC. For WC, a setting of $\rho_{WC} = 0.97$ always returns a bootstopped set with a WRF $< 2\%$ to the MRE consensus of the reference replicates. The results also clearly show that there is a significant alignment-dependent variability in the stoppping numbers, as these range between 150 and 450 replicates for FC and between 50 and 1,200 for WC.

In Table 2 we provide additional metrics for the bootstopped trees. Columns μ_x and σ_x^2 provide the mean error and the mean squared error between support values induced by the $x = \{$FC,WC$\}$-bootstopped trees and by the reference trees on the best-scoring ML tree. Columns *SUPPLOSS-FC* and *SUPPLOSS-WC* quantify the deviations of support values in the best scoring ML tree.

In Figure 2 we graphically depict, for one dataset (1481), the convergence of FC versus WC. We plot the RF and WRF distances between the MRE consensus of the bootstopped trees and reference trees over distinct settings (0.87, 0.88,. . . ,0.99)

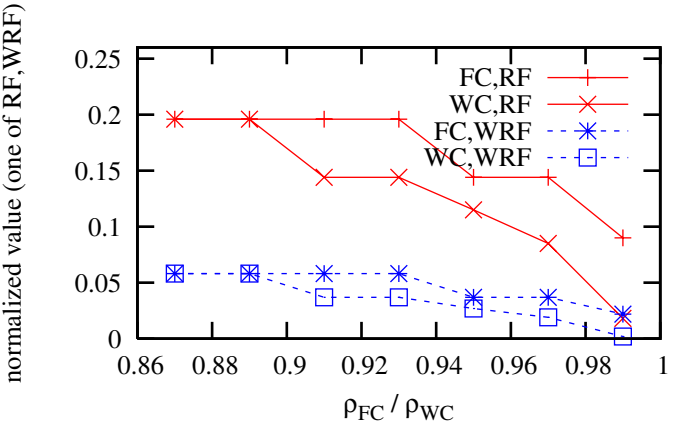

Fig. 2. Plot showing convergence of WC over FC for various threshold settings (ρ_{FC} and ρ_{WC} respectively) on dataset 1418

for ρ_{FC} and ρ_{WC}. For all but two datasets we observed that WC yielded a better convergence (while it required almost 50% more repliocates on average) toward replicate sets whose consensi are more congruent (i.e., have lower RF and WRF distances) with the full replicate sets, as a function of ρ. This favorable property is due to the fact that WC is exclusively based on the "important" bipartitions. Therefore, WC allows to more precisely specify the desired degree of accuracy with respect to the biologically relevant information via an appropriate setting of ρ. As can be derived from Table 1 a setting of $\rho = 0.97$ for WC induces a WRF toward the reference dataset consensus that is $\leq 2\%$ in all cases for all of our datasets. Hence, the usage of a WC threshold will also be more meaningful, because it appears to be strongly correlated with the final WRF distance to the 10,000 reference replicates.

In addition to assessing our stopping criteria, we have also comprehensively assessed the inherent convergence properties of our replicate sets. Doing so has enabled us to understand a number of quantities that tend to reflect bootstrap support and may help in the design of improved stopping criteria. We have plotted a number of (dis)similarity measures between a subset (i.e., the first m trees) and full replicate ($\geq 10,000$ trees) set. In Figure 3 we plot the RF and WRF (in the two lower plots) between the MRE consensus of each tree set restricted to the first m trees versus its respective full set of replicates ($\geq 10,000$ trees). This plot shows the differences in convergence speeds among datasets. In addition, it underlines that WRF introduces less noise than RF as replicates are added, so that WRF is a more reliable measure for convergence. An extreme example for this is dataset 354, a short (348 alignment columns) alignment of maple tree sequences from the ITS gene that is known to be hard to analyze [11]. A comparison between the development of RF and WRF over the number of trees for this alignment shows that there are many sequences with low support that are placed in different parts of the tree and essentially reflect unresolved nodes. The slight increase of distance metrics around 1,000 replicates and consecutive decrease observed for dataset 125 might be minor artifacts of the RAxML RBS algorithm.

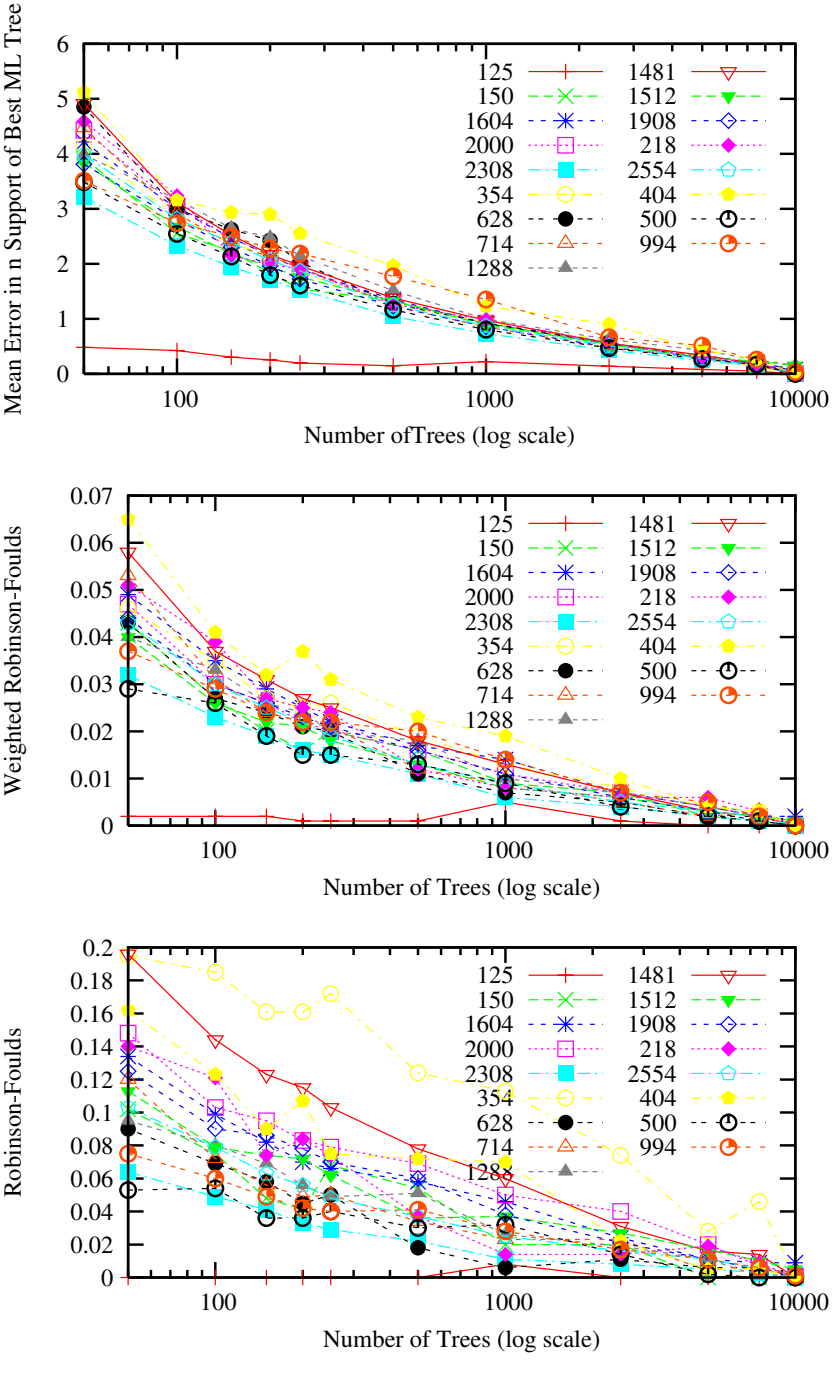

Fig. 3. Inherent convergence of replicate sets scored by (top) error in support of best ML tree (middle) WRF and (bottom) RF distances between the first m trees and the entire (10,000 tree) set

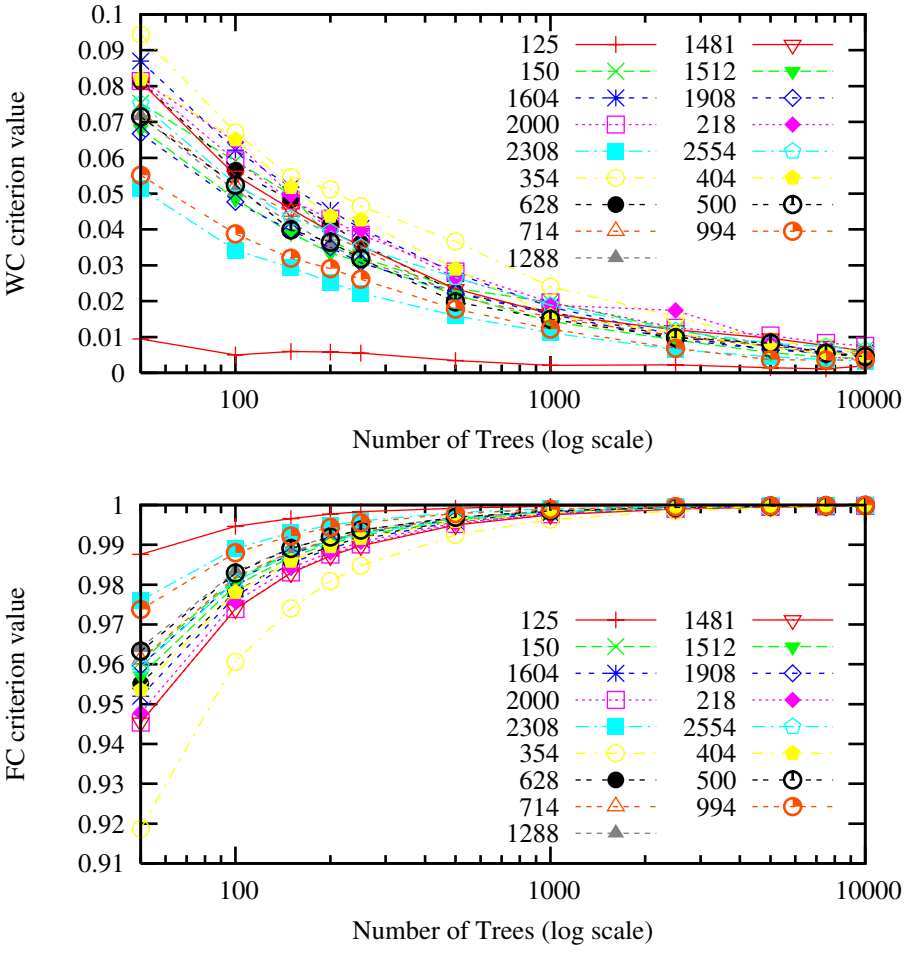

Fig. 4. Values of FC and WC criteria for tree subsets consisting of the first m trees

Also in the upper part of Figure 3 we plot the development of the mean error between support values of m replicates and all replicates on the best-scoring tree. The three plots in Figure 3 clearly show that the development of WRF distances over the number of replicates is highly congruent to the development of the mean error on the best-scoring tree. Thus, WRF can be used as a criterion to determine convergence without an external reference tree. Designing such a criterion has been a major goal of the phylogenetic community; WRF is the first good answer. Moreover, the plots can help to determine an appropriate threshold setting for ρ_{WC}, depending on the desired degree of accuracy.

Finally, in Figure 5 we plot the support values of FC/WC-bootstopped trees against the support values from the reference replicates on the best-scoring ML tree for dataset 628. The comparison clearly shows a decrease in deviations from the diagonal for the WC criterion.

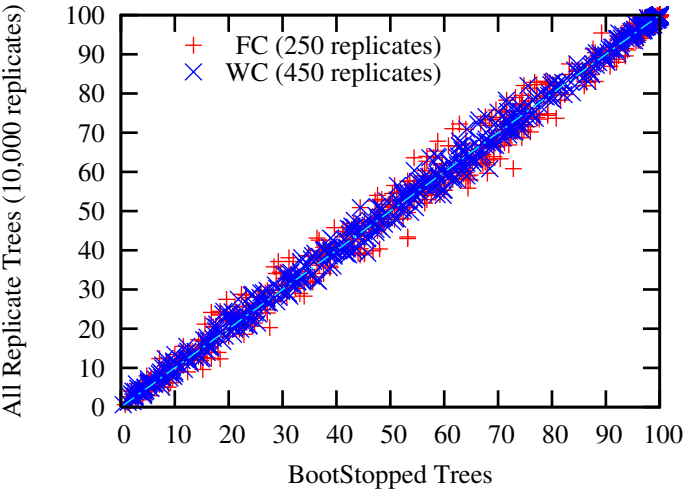

Fig. 5. Support values drawn on the best ML tree for FC (blue) and WC (red) versus full replicate set, for data set 628

5 Conclusion and Future Work

We have conducted the first large-scale empirical ML-based study of the convergence properties of bootstrapping, using biological datasets that cover a wide range of input alignment sizes and a broad variety of organisms and genes. In addition, we have developped and assessed two bootstopping criteria that can be computed at run time and do not rely on externally provided reference trees to determine convergence. The criteria have been designed so as to capture a stopping point that provides sufficient accuracy for an unambigoous biological interpretation of the resulting consensus trees or best-known ML trees with support values. The correlation between bootstopped support values and support values from 10,000 reference trees exceeds 99.5% in all cases, while the relative weighted tree distance (used with the WC criterion) is smaller than the specified threshold value in all cases. We conclude that the WC criterion yields better performance and higher accuracy than FC while it correlates very well with the mean error of support values on the best-scoring tree. We advocate the use of WC over FC because it only takes into account the BS support of "important" bipartitions which are subject to biological interpretation. We have also shown that the number of replicates required to achieve a certain level of accuracy is highly dataset-dependent for real data, so that, by using our criteria, an investigator need only compute as many replicates as necessary, thus avoiding the waste of scarce computational resources, in particular for future large-scale phylogenomic analyses.

Future work entails the full integration of the bootstopping criteria into the forthcoming release of RAxML (RAxML 7.1.0). We will assess whether performance gains can be obtained by applying embedding techniques in the

calculation of RF/WRF[22]. We will also devise ways to dynamically adapt the spacing of FC/WC criteria (which is currently fixed at 50) to the convergence speed of the BS replicates, i.e., use a more sparse spacing for the initial phase and a denser spacing for the later phase of the BS search.

Acknowledgments

We would like to thank Derrick Zwickl and Bret Larget for useful discussions on this manuscript. We are also thankful to Andrew Rambaut for discussions on Trace and AWTY. We would also like to thank the following colleagues for providing real-world datasets: N. Poulakakis, U. Roshan, M. Gottschling, M. Göker, G. Grimm, C. Robertson, N. Salamin. Part of this work was funded under the auspices of the Emmy Noether program by the German Science Foundation (DFG).

References

1. Andrews, D.W.K., Buchinsky, M.: On the Number of Bootstrap Repetitions for Bootstrap Standard Errors, Confidence Intervals, and Tests. Cowles Foundation Paper 1141R (1997)
2. Andrews, D.W.K., Buchinsky, M.: A Three-Step Method for Choosing the Number of Bootstrap Repetitions. Econometrica 68(1), 23–51 (2000)
3. Andrews, D.W.K., Buchinsky, M.: Evaluation of a Three-step Method for Choosing the Number of Bootstrap Repetitions. J. of Econometrics 103(1-2), 345–386 (2001)
4. Andrews, D.W.K., Buchinsky, M.: On The Number of Bootstrap Repetitions for BCa Confidence Intervals. Econometric Theory 18(4), 962–984 (2002)
5. Brooks, S.P., Gelman, A.: General Methods for Monitoring Convergence of Iterative Simulations. J. of Computational and Graphical Statistics 7(4), 434–455 (1998)
6. Davidson, A.C., Hinkley, D.V.: Bootstrap Methods and Their Application. Cambridge University Press, Cambridge (2003)
7. Davidson, R., MacKinnon, J.G.: Bootstrap Tests: How Many Bootstraps? Econometric Reviews 19(1), 55–68 (2000)
8. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman and Hall, New York (1993)
9. Felsenstein, J.: Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Evolution 39(4), 783–791 (1985)
10. Gelman, A., Rubin, D.B.: Inference from Iterative Simulation using Multiple Sequences. Stat. Sci. 7, 457–511 (1992)
11. Grimm, G.W., Renner, S.S., Stamatakis, A., Hemleben, V.: A Nuclear Ribosomal DNA Phylogeny of acer Inferred with Maximum Likelihood, Splits Graphs, and Motif Analyses of 606 Sequences. Evolutionary Bioinformatics Online 2, 279–294 (2006)
12. Guindon, S., Gascuel, O.: A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. Sys. Biol. 52(5), 696–704 (2003)
13. Guo, W., Peddada, S.: Adaptive Choice of the Number of Bootstrap Samples in Large Scale Multiple Testing. Stat. Appls. in Genetics and Mol. Biol. 7(1) (2008)
14. Hall, P.: On the Number of Bootstrap Simulations Required to Construct a Confidence Interval. The Annals of Statistics 14(4), 1453–1462 (1986)

15. Hedges, S.B.: The Number of Replications Needed for Accurate Estimation of the Bootstrap P Value in Phylogenetic Studies. Mol. Biol. Evol. 9(2), 366–369 (1992)
16. Hillis, D.M., Heath, T.A., John, K.S.: Analysis and Visualization of Tree Space. Sys. Biol. 54(3), 471–482 (2005)
17. Holmes, S.: Bootstrapping Phylogenies Statistical Science, 18(2), 241–255
18. Manly, B.F.J., et al.: Randomization, Bootstrap and Monte Carlo Methods in Biology. CRC Press, Boca Raton (1997)
19. Moret, B.M.E.: Large-scale Phylogenetic Reconstruction. In: Brown, J.R. (ed.) Comparative Genomics: Basic and Applied Research, pp. 29–48. CRC Press/Taylor & Francis (2007)
20. Mossel, E., Vigoda, E.: Limitations of Markov Chain Monte Carlo Algorithms for Bayesian Inference of Phylogeny. Ann. Appl. Probab. 16(4), 2215–2234 (2006)
21. Nylander, J.A.A., Wilgenbusch, J.C., Warren, D.L., Swofford, D.L.: AWTY (are we there yet?): A System for Graphical Exploration of MCMC Convergence in Bayesian Phylogenetics. Bioinformatics (2007) (advance access, published August 30)
22. Pattengale, N.D., Gottlieb, E.J., Moret, B.M.E.: Efficiently Computing the Robinson-Foulds Metric. J. of Computational Biology 14(6), 724–735 (2007)
23. Rambaut, A., Drummond, A.: Tracer MCMC Trace Analysis Tool version 1.3 (2004)
24. Robinson, D.F., Foulds, L.R.: Comparison of Weighted Labelled Trees. Lecture Notes in Mathematics 748, 119–126 (1979)
25. Robinson, D.F., Foulds, L.R.: Comparison of Phylogenetic Trees. Math. Biosc. 53(1), 131–147 (1981)
26. Ronquist, F., Huelsenbeck, J.P.: MrBayes 3: Bayesian Phylogenetic Inference under Mixed Models. Bioinformatics 19(12), 1572–1574 (2003)
27. Soltis, D.E., Gitzendanner, M.A., Soltis, P.S.: A 567-taxon Data Set for Angiosperms: The Challenges Posed by Bayesian Analyses of Large Data Sets. Int'l J. Plant Sci. 168(2), 137–157 (2007)
28. Soltis, D.E., Soltis, P.S.: Applying the Bootstrap in Phylogeny Reconstruction. Statist. Sci. 18(2), 256–267 (2003)
29. Stamatakis, A.: RAxML-VI-HPC: Maximum Likelihood-based Phylogenetic Analyses with Thousands of Taxa and Mixed Models. Bioinformatics 22(21), 2688–2690 (2006)
30. Stamatakis, A., Hoover, P., Rougemont, J.: A Rapid Bootstrap Algorithm for the RAxML Web Servers. Sys. Biol. (2008) (in press)
31. Stamatakis, A., Meier, H., Ludwig, T.: New Fast and Accurate Heuristics for Inference of Large Phylogenetic Trees. In: Proc. of IPDPS 2004, HICOMB Workshop, Proceedings on CD, Santa Fe, New Mexico (2004)
32. Whelan, S.: New Approaches to Phylogenetic Tree Search and Their Application to Large Numbers of Protein Alignments. Sys. Biol. 56(5), 727–740 (2007)
33. Zwickl, D.: Genetic Algorithm Approaches for the Phylogenetic Analysis of Large Biological Sequence Datasets under the Maximum Likelihood Criterion. PhD thesis, University of Texas at Austin (April 2006)