A Simple and Accurate Method for Rogue Taxon Identification

Andre J. Aberer
Scientific Computing Group
Heidelberg Institute for Theoretical Studies,
Heidelberg, Germany
andre.aberer@h-its.org

Alexandros Stamatakis
Scientific Computing Group
Heidelberg Institute for Theoretical Studies
Heidelberg, Germany
alexandros.stamatakis@h-its.org

***Abstract*—The summary of a phylogenetic analysis (typically a consensus tree) can be substantially biased by so-called *rogue taxa* (or briefly: rogues). Rogue taxa assume varying phylogenetic positions in the tree collection that is used to build the consensus tree and thereby decrease the resolution of the consensus.**

We present an accurate and straight-forward algorithm for identifying rogues that assesses the effect on the consensus tree support values by removing one taxon at a time. Our approach improves the resolution of the consensus tree *and*, at the same time, increases the support values of existing relationships. We compare our algorithm to three competing methods (leaf stability index, taxonomic instability index, and Pattengale's algorithm) on a large number of real biological data sets. We show that it outperforms stability-based methods since rogue taxa are not necessarily the most unstable taxa with respect to stability measures. Our algorithm is substantially more memory-efficient than Pattengale's approach while instances, where Pattengale's algorithm outperforms our approach, appear to be rare on real data. Finally, we find that, it is advisable to conduct a *de novo* bootstrap analysis after rogues have been removed from the sequence alignment.

***Keywords*-rogue taxa; consensus tree; phylogenetic post-analysis; taxonomic instability index; leaf stability index;**

I. Introduction

The central goal in phylogenetics is to disentangle the evolutionary history of species (also referred to as *taxa*) given genetic information (typically provided as multiple sequence alignment (MSA)). Phylogenies are represented as unrooted binary trees. The leaves (degree-1-nodes) represent the taxa under study, while inner nodes represent hypothetical common ancestors. Two common criteria for inferring phylogenies are maximum likelihood (ML) or maximum parsimony (MP) (reviewed in [1]). Tree searches for the optimal tree on the original MSA under ML and MP are usually supplemented by a non-parametric bootstrap analysis [2] to obtain support values for specific evolutionary relationships. The bootstrap procedure randomly draws columns/sites with replacement from the original MSA until a MSA replicate with the same number of sites (but a different site composition) as the original MSA has been assembled. Several MSA bootstrap replicates are generated and a tree is inferred on each replicate. The information contained in such a set of bootstrap trees can then be drawn onto the best-known (ML and MP are $\mathcal{NP}$-hard [3], [4]) tree as branch support values. Alternatively, the bootstrap tree set can be summarized by a consensus tree.

To compute a consensus tree, the set of bootstrap trees is initially transformed into a corresponding *bipartition* list. A bipartition is obtained by removing an edge that connects two inner nodes from a tree. This yields two disjoint partitions A and $\overline{A}$, that is, a bipartition of the entire taxon set. There exist several methods (reviewed in [5]) for building consensus trees from a bipartition list. Here, we focus on the widely-used majority-rule consensus (MRC) and strict consensus (SC) methods. MRC builds a consensus tree from those bipartitions that occur in at least half of the trees under consideration, while a strict consensus (SC) tree only contains those bipartitions that occur in all trees. In Fig. 1 we provide a simple example of MRC and SC consensus trees for a tree set of size two. Because the two trees in the tree set do not share a bipartition, SC and MRC yield a completely unresolved star tree (see Fig. 1(c)).

The lack of resolution in the consensi is caused by the two rogue taxa R and Q that assume different positions in the input tree set. Wilkinson first examined such wandering taxa and coined the term *rogue taxon* [6]–[8]. The presence of just a small number of rogue taxa (also denoted as rogues) in the tree set (typically bootstrap replicate trees) can either yield a completely unresolved star tree (see Fig. 1(c)) or substantially decrease the accumulated bipartition support in the consensus. If we are able to identify and prune (remove) the two rogue taxa R and Q from the input trees in our example, we can recover shared bipartitions in the input trees and thereby obtain a more informative consensus tree for the remaining taxa (see Fig. 1(d)).

Rogue taxa may occur for various reasons: (i) general lack of phylogenetic signal (e.g., because of an excessive proportion gaps in the alignment, or either too high or too low mutation rates [9]); (ii) ambiguous phylogenetic signal because of mislabeled or erroneous sequences (specifically *chimeric sequences*) or horizontal gene transfer. Rogues can also occur in super-matrices (when genes that only entail a subset of taxa are concatenated) [10]. Rogue taxa also seem to be problematic when super-tree approaches [11] are deployed (an alternative approach to super-matrices for

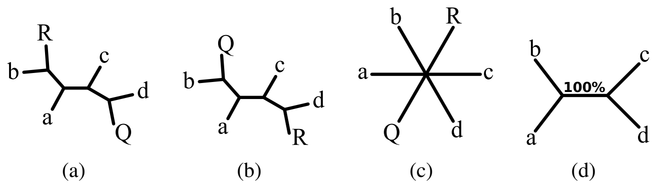


Figure 1: a) and b) two bootstrap trees with 6 taxa. c) the strict and majority-rule consensus tree of the two bootstrap trees. d) both consensus trees contain a bipartition after rogue taxa R and Q have been pruned.

harnessing huge amounts of data).

While, biologists may be able to identify rogue taxa by visual inspection in small data sets, the general trend for larger phylogenies to proportionally contain more rogues (see Section IV-A) underlines the need for automated rogue taxon identification. Currently, only a few phylogenetic studies include a systematic rogue taxon analysis [10], [12]–[14]. Typically, the topological stability for each taxon is computed and the most unstable taxa are pruned from the bootstrap trees. Per-taxon topological stability is quantified by the leaf stability index [15] or the taxonomic instability value as implemented in Mesquite [16].

Here we introduce and make available a new, simple, yet accurate method for rogue taxon identification. We also conduct an extensive comparison of our rogue taxon identification method with competing approaches on a large and diverse collection of real biological data sets. We also address the question, if pruning identified rogues from an existing set of bootstrap replicate trees is sufficient for obtaining an accurate consensus tree, or whether it is necessary to re-run the bootstrap analysis from scratch on the pruned MSA.

Note that, the methods we analyze here, remove identified rogues from bootstrap tree replicates with the aim to improve support values in consensus trees. This does not reflect the way biologists would ideally like to handle rogues, since they are interested in the phylogenetic relationships among all taxa in the initial MSA. Nonetheless, this represents a first step toward adressing the problem.

The remainder of this paper is organized as follows. Initially, we discuss related approaches for rogue taxon identification in Section II. In Section III, we introduce our rogue taxon identification algorithm. In Section IV we compare the performance of our algorithm to alternative rogue identification methods and assess if re-running the bootstrap analysis after rogue identification and removal is required. We conclude and discuss directions of future work in Section V.

II. Related Work

We briefly cover two related approaches for rogue identification that rely on the idea that rogues are the most topologically unstable taxa in a set of bootstrap trees.

A. Node Stability Measures

The *taxonomic instability index* (TII) [16] captures taxon stability by means of the variation of pair-wise distances between taxon pairs across all bootstrap trees. The taxonomic instability index of a taxon i is defined as $\sum_{(x,y)} \frac{|d_{ijx}-d_{ijy}|}{(d_{ijx}+d_{ijy})^z}$, where d_{ijx} is the unweighted patristic distance between two non-identical taxa i and j (i.e., the number of nodes on the path between i and j in bootstrap tree x). The z parameter is used to specify whether evolutionary distant (in case of large values of z) or evolutionary close relationships (small values of z) have a higher impact on the TII. The default value in the Mesquite implementation is $z := 2$. This represents a natural choice for rogue identification, since we expect rogues to assume varying positions that are nonetheless located close to each other. A taxon is interpreted as stable (i.e., non-rogue) with respect to the TII, when the unweighted patristic distance to other taxa remains relatively constant over all bootstrap trees.

The *leaf stability index* (LSI) [15] uses taxon triplets. In a *rooted* binary tree, there exist three different possibilities for the relationship of three taxa: a, b, c. The three possible triplets (relationships) are $((a,b),c)$, $((a,c),b)$ and $((b,c),a)$. The triplet stability of a taxon triplet a, b, c is defined as the difference of the relative frequency of the most prevalent triplet and the second most prevalent triplet in the bootstrap trees. The LSI for a taxon a is defined as the average triplet stability over all triplets that contain a.

The LSI is normalized to $[0.0...1.0]$. In contrast to the TII, this allows to use absolute cut-off values. Since relationship of a stable taxon to all other taxa in the bootstrap trees will not vary considerably, such a stable taxon will occur predominantly in stable triplets. Hence, the LSI of a stable taxon is close to the maximum LSI value of 1.0. Relationships to evolutionary distant taxa have the same impact on the LSI as evolutionary close relationships, whereas for the TII this trade-off can be adjusted by the z parameter.

The LSI as used here represents one of the four variants of this stability measure and is the most frequently used variant (the results of alternative LSI variants are mostly equivalent; see supplementary material of [13]).

While the TII can be applied as-is to unrooted trees, some minor adaptations are required for the LSI (see [17]) by considering quartets of taxa in unrooted trees rather than triplets. The LSI has been used in some recent phylogenetic studies [12], [13].

LSI and TII identify taxa R and Q in our example in Fig. 1 as the two most unstable taxa. Thus, if the somewhat arbitrary cutoff threshold for pruning taxa is chosen adequately, R and Q will be correctly identified as rogues.

B. Merging Low-Support Bipartitions

Recently, Pattengale [18] proposed a method to directly optimize the number of bipartitions in a consensus tree by identifying and removing rogues. In most cases, pruning a set of taxa from the initial tree set, will alter the number of bipartitions included in the respective consensus tree. This is because, pruning taxa can merge some bipartitions with each other, while other bipartitions may vanish (when the number of taxa in one of the partitions becomes ≤ 1). As an example for merging bipartitions consider the bipartitions $\mathbf{A} = A|\overline{A} = abR|cdQ$ in Fig. 1(a) and $\mathbf{B} = B|\overline{B} = abQ|cdR$ in Fig. 1(b), where a, b, c, d, Q, R are the respective taxon names. If we prune rogue taxa R and Q from the trees in Fig. 1, the two bipartitions $\mathbf{A}$ (A and $\overline{A}$) and $\mathbf{B}$ (A and $\overline{A}$) become identical and can be merged into a single bipartition $\mathbf{C} = ab|cd$. Bipartition $\mathbf{C}$ is now contained in both pruned input trees and therefore defines an inner branch in the SC and MRC consensi.

The above observation is the basis of Pattengale's algorithm (henceforth denoted as *bipartition merging algorithm* (BMA)) that works as follows: At an abstract level, the algorithm strives to determine a set of taxa (a so-called *drop set*) that, if pruned, will yield a maximum increase of bipartitions in the consensus tree.

To determine the drop set, the algorithm initially computes a *bipartition profile*, which maps each bipartition of the tree set to the set of trees in which it occurs. For each pair of bipartitions in this profile that is not already part of the consensus tree, the algorithm determines the required drop set (taxa to be pruned) that will merge the bipartition pair. If the resulting merged bipartition is contained in the consensus tree, the corresponding drop set is added to a data structure that maps drop sets to a list of bipartition merging events.

This drop set mapping is then used to select the drop set (and prune the taxa therein) that will add the maximum number of bipartitions to the consensus tree *and* requires pruning the minimum number taxa from the trees. Both criteria (minimize number of taxa to prune *and* maximize number of consensus tree bipartitions added) are incorporated into a single measure for optimizing this bicriterion problem called the *relative information content* (RIC). The RIC of a consensus tree C' for a pruned bootstrap tree collection is defined as $\text{RIC}(C') = \frac{B'+T'}{T-3+T}$, where B' is the number of bipartitions in C' and T' is the number of remaining taxa after pruning. The denominator normalizes the RIC by the number of taxa in the initial taxon set T and the maximum number $T-3$ of possible consensus bipartitions in the initial consensus tree C.

The above procedure for identifying (and pruning) rogue taxa is applied iteratively until the RIC measure can not be further improved. In other words, the algorithm will only prune drop sets where the number of taxa pruned is smaller than the number of bipartitions added to the consensus tree by pruning those taxa.

The BMA allows for rapid identification of rogue taxa in large data sets. However, it only represents a greedy approximation, since the BMA does not check if existing consensus bipartitions vanish or merge when a drop set is pruned. Thus, the BMA may prune sub-optimal drop sets. Furthermore, the algorithm can not detect potential merging events that occur if more than two bipartitions are merged into a consensus bipartition. The algorithm will also yield an approximation error, when *subsets* of a drop set give rise to additional consensus tree bipartitions.

III. Rogue Taxon Identification Algorithm

Initially, we formulate an optimization criterion that is similar to that of Pattengale *et al.* and propose a simple approximation algorithm. Then, we describe the experimental setup for assessing alternative rogue identification methods.

A. Optimality Criterion

Apart from the number of remaining taxa, the RIC criterion incorporates the number of bipartitions in the consensus tree. In our optimality criterion, we opt for a more fine-grain measure. Let $\text{sup}(\mathbf{B_i})$ be the relative frequency of a bipartition $\mathbf{B_i}$ in the bootstrap trees, l the number of consensus bipartitions, and T the number of taxa in the initial taxon set. Our optimality criterion, the *relative bipartition information content* (RBIC) of a consensus tree C' is defined as

$$\text{RBIC}(C') = \frac{\sum_{i=\{1..l\}} \text{sup}(\mathbf{B_i})}{T-3}.$$

Using the relative frequencies (i.e., support or frequency of occurrence) of bipartitions for the RBIC has the advantage that optimizing this criterion will preferably recover highly supported bipartitions. In the most extreme case, the algorithm can choose between pruning a taxon a for recovering a bipartition $\mathbf{B_i}$ with $\text{sup}(\mathbf{B_i}) = 50\%$ or pruning a taxon b for recovering a bipartition $\mathbf{B_j}$ with $\text{sup}(\mathbf{B_j}) = 100\%$. Using the RBIC will not only lead to pruning taxa that add bipartitions to the consensus tree, but also prune taxa that improve the support of existing consensus bipartitions.

The RBIC is normalized by the maximum possible support in a fully resolved consensus tree prior to pruning any rogues. Thereby, the RBIC of a consensus tree C' represents the (attained) proportion of the maximum possible support in a consensus tree for the initial set of taxa. We did not include the term T' in the RBIC (in contrast to the RIC of Pattengale *et al.*). Instead, we optimize for maximal accumulated support and let the user decide if, at some point, improvements are too small to justify pruning additional taxa.

B. Algorithm Description

The BMA does not fit the RBIC criterion particularly well. According to initial experiments, the aforementioned approximation errors have a more pronounced effect on the support-based RBIC criterion than on the bipartition-based RIC criterion. Furthermore, runtime requirements of the BMA would increase significantly, since some RIC-specific optimizations we can not be applied.

We therefore decided to directly calculate consensus trees from sets of pruned bootstrap trees. The only (yet significant) simplification in our algorithm is that, as opposed to the BMA, we only test the impact of pruning one taxon at a time to assess the corresponding RBIC improvement. The taxon that yields the highest RBIC improvement (the most "rogue" taxon) is then permanently removed from the bootstrap trees.

Once the most "rogue" taxon has been pruned, we iteratively apply this algorithm to the pruned tree set. In each iteration we re-calculate the RBIC change induced by pruning all remaining taxa (one at a time) and remove the next most "rogue" taxon permanently. We repeat this procedure until the RBIC can not be further improved. We refer to this algorithm as *single-taxon algorithm* (STA).

The STA is the only rogue identification method that can not optimally solve the initial example in Fig. 1. The only way to improve the RBIC of the consensus in this example is to prune rogue taxa R and Q simultaneously. Thus, the STA will not obtain an RBIC improvement by just pruning one of them and terminates without pruning a taxon. However, this example has been constructed to demonstrate the algorithm's worst-case behavior. Evidence that such configurations are uncommon in real biological data is presented in Section IV-B.

The STA implementation relies on the optimized consensus tree algorithm implementations in RAxML [19]. We compute the bipartition profile of the tree set only once at program initialization. Then, we compute the RBIC-change induced by pruning one taxon at a time over all taxa in parallel. Thus, the STA can be applied to comparatively large data sets in terms of number of trees and/or number of taxa (see Section III-C). Compared to the highly optimized implementation of the BMA [18], the STA implementation is up to two orders of magnitude slower (per iteration and for data sets up to 714 taxa). However, the memory requirements of STA are constant and proportional to the size of the bipartition profile n, whereas the space requirements of BMA are in $\mathcal{O}(n^2)$ which limits its scalability [18].

C. Experimental Setup

For our comparative analysis we used a bootstrap tree collection from a previous study by Pattengale [20] (some of these data sets have also been used in the study introducing the BMA [18]). Furthermore, the RAxML user community provided rogue-suspicious data sets. We received 17 real-world MSAs from the users and computed 1,000 bootstrap trees on each using the RAxML rapid bootstrap algorithm (RAxML v7.2.8 [21]) under the GTR+CAT approximation [22] of rate heterogeneity for DNA alignments and using Dayhoff+CAT for amino acid alignments. The arbitrarily chosen protein substitution model is of minor importance because the key goal was to generate a sufficiently large number of bootstrap replicates. The number of taxa in the data sets ranges from 24 up to 2,254 taxa (see Table I). Since some data sets are still unpublished, not all are available for download yet at www.exelixis-lab.org/software/data.tbz.

The alternative rogue identification methods (TII, LSI, BMA) were also integrated into the RAxML framework (available at www.exelixis-lab.org/software/raxml-rogue-edition.tbz). For TII *and* LSI we implemented two variants: a *monolithic* version that initially computes the stability measure and subsequently prunes a given number of taxa as well as an *iterative* version. The iterative version determines *the* most unstable taxon *de novo* at each iteration and then immediately prunes it prior to recomputing the stability scores on the reduced taxon set in analogy to the STA. For our comparative analysis (see Section IV-A), we pruned up to one third of the initial taxon set. For each of the above rogue identification methods, we selected the (intermediate) pruning step that maximized the RBIC criterion, as the optimal solution.

IV. Results

A. Performance of the Single Taxon Algorithm

We initially tested performance of STA, LSI, and TII (see Section III-C; see below for an assessment of BMA). For each data set, Table I shows the RBIC of the respective pruned consensus trees for the three alternative methods and the number of taxa that were pruned. The effect of the intermediate pruning steps on the RBIC is depicted in Fig. 2 for the data sets with the highest improvement potential.

In almost all cases, STA achieves a higher RBIC in the pruned consensus tree than the competing methods (see Table I). On 12 data sets the RBIC improvement obtained by STA is better by at least one order of magnitude than the RBIC improvement achieved by LSI or TII. For instance, on the MSA with 1,481 taxa, STA improves the RBIC from 0.438 to 0.514, while TII achieves a RBIC score of 0.440. Similarly, for the 404-taxon MSA (see Fig. 2(c)), the STA method increases the RBIC from 0.500 to 0.609, while TII attains an RBIC of 0.539 and LSI yields an RBIC of 0.515. STA also performs particularly well on the MSAs with 72 (Fig. 2(a)), 128 (Fig. 2(b)), and 885 (Fig. 2(d)) taxa. Note that, for the 885 taxon data set (Fig. 2(d)), the initial consensus is particularly weak (RBIC: 0.162). The STA prunes 59% of the taxa until the maximum RBIC is attained. Most of these rogues have a weak impact on the RBIC. As the MSA consists of only 623 characters/sites, we assume that an overall lack of phylogenetic signal may be the reason for this behavior.

The iterative versions of the stability indices do not perform consistently better than their monolithic counterparts (see Table I). Thus, the specific implementation of the stability index methods (iterative versus monolithic) does not affect their accuracy.

Except for the 24- and 628-taxon data sets, the TII always yields results that are at least as good as those obtained by the LSI (see Table I). Remember that, for the LSI, all

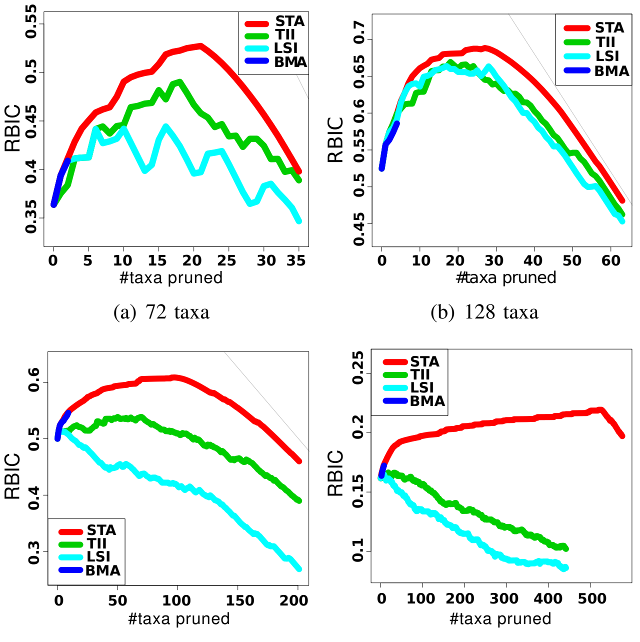
#taxa	initial RBIC	bip. merging algorithm		single-taxon algorithm		taxonomic instab. index				leaf stability index			
						monolithic		iterative		monolithic		iterative	
		#p	RBIC	#p	RBIC	#p	RBIC	#p	RBIC	#p	RBIC	#p	RBIC
24	0.829	0	0.829	1	0.853	3	0.837	3	0.837	1	0.853	1	0.853
44	0.755	1	0.786	2	0.789	2	0.786	2	0.786	1	0.786	1	0.786
52	0.543	1	0.582	9	0.653	6	0.639	6	0.627	7	0.622	7	0.622
72	0.364	2	0.409	21	0.527	18	0.490	19	0.496	16	0.444	7	0.444
88	0.783	1	0.810	5	0.832	4	0.832	4	0.832	5	0.813	2	0.813
125	0.958	0	0.958	1	0.959	0	0.958	0	0.958	0	0.958	0	0.958
128	0.525	4	0.586	27	0.688	18	0.669	21	0.667	17	0.664	17	0.657
141	0.669	2	0.690	15	0.709	11	0.706	11	0.706	2	0.686	2	0.686
143	0.610	2	0.628	12	0.651	3	0.628	20	0.632	7	0.619	7	0.619
148	0.611	2	0.647	17	0.672	10	0.651	7	0.655	3	0.647	3	0.647
150	0.570	1	0.579	28	0.614	14	0.581	17	0.589	6	0.571	5	0.571
218	0.471	4	0.485	51	0.548	4	0.481	4	0.481	3	0.473	3	0.473
316	0.365	2	0.373	113	0.438	11	0.371	11	0.371	0	0.365	0	0.365
317	0.541	3	0.551	75	0.610	6	0.552	6	0.549	0	0.541	0	0.541
350	0.495	4	0.508	62	0.555	9	0.497	5	0.497	0	0.495	0	0.495
354	0.328	2	0.334	146	0.386	3	0.328	3	0.328	2	0.328	0	0.328
404	0.500	11	0.548	97	0.609	50	0.539	63	0.549	1	0.515		
424	0.501	11	0.537	80	0.607	55	0.540	46	0.541	13	0.508		
451	0.492	12	0.529	100	0.601	36	0.538	35	0.541	30	0.508		
500	0.589	4	0.597	92	0.634	28	0.598	22	0.597	7	0.594		
628	0.515	4	0.525	133	0.564	4	0.516	13	0.515	3	0.520		
714	0.569	6	0.579	125	0.620	6	0.570	6	0.570	2	0.569		
885	0.162	8	0.172	525	0.219	18	0.167	36	0.167	6	0.165		
994	0.679	3	0.686	95	0.704	2	0.681	26	0.682	1	0.681		
1,288	0.584	7	0.591	217	0.634	13	0.587						
1,481	0.438	19	0.454	428	0.514	3	0.440						
1,512	0.518	11	0.530	347	0.568	6	0.524						
1,604	0.487	19	0.502	396	0.548	14	0.490						
1,908	0.553	11	0.557	383	0.590	0	0.553						
2,000	0.449	40	0.479	479	0.526	42	0.456						
2,308	0.713	9	0.718	196	0.734	9	0.716						
2,554	0.541	37	0.555	523	0.596	51	0.543						

Table I: Performance comparison: RBIC of the unpruned consensus trees of all data sets under study compared to the RBIC maximum achieved by pruning taxa according to the respective algorithms. Some data points are missing for some algorithms because of prohibitive execution times. For each method, the maximum achieved RBIC (**RBIC**) and number of pruned taxa at the maximum (**#p**) is specified.

stability relationships (in form of quartet frequencies) have the same impact on the measure. Thus, the LSI strives to measure taxon stability at a *global* level, while—depending on the z value—TII can identify taxon stability at a more local level (see Section II-A). Thus, the TII predominantly captures the *local* stability of a taxon. In this context, the comparison between STA and the LSI/TII implies that rogue taxa may be unstable at a local and (although less likely) global scale. Inversely, unstable taxa are not necessarily rogue taxa. This hypothesis is underlined by the observation that, the performance gap between the three methods increases with data set size.

Fig. 2 also provides performance data for BMA. We can not directly compare the maxima of the BMA curves to STA performance, since the two approaches intend to answer different questions. There exist three reasons why the two performance curves can potentially disagree until BMA convergence: (i) the data set contains interdependent rogue taxa (as in our initial example) that can not be identified by assessing the effect of pruning only one taxon at a time, (ii) the approximation in the BMA prunes a sub-optimal set of taxa, and (iii) the STA favors a taxon that increases support in the consensus tree but does not generate a new bipartition (STA optimizes the more fine-grained RBIC instead of the RIC). However, until convergence of the BMA, both algorithms choose largely identical taxa. Overall, only 15.6% of taxa that are pruned by BMA are not pruned by STA. Thus, the observed experimental agreement between both approaches indicates that the aforementioned phenomena (i), (ii), and (iii) that can lead to disagreements do not occur often.

Given the lack of potential disagreement, we compare the maximum RBIC values achieved by the BMA and the STA. We observe that, the STA achieves an RBIC improvement that is up to 8.7 times higher (average: 3.4) than the RBIC improvement attained by BMA, while it prunes up to 72 times more taxa (average: 20.7). Thus, in general, the less conservative stopping criterion in STA combined with the more fine-grain RBIC optimality criterion can substantially increase consensus support. Nevertheless, some data sets (e.g., Fig. 2(c) and 2(d)) contain a large number of taxa that only induce a weak RBIC improvement. On such data sets, users may thus opt to stop pruning taxa before the RBIC maximum is reached.



(a) 72 taxa

(b) 128 taxa

(c) 404 taxa

(d) 885 taxa

Figure 2: Performance comparison: RBIC optimality of intermediate pruning steps of STA, the two (monolithic) stability index algorithms and BMA on 4 data sets. Gray line (where present) represents the maximum RBIC that can be achieved after the given number of taxa have been pruned.

B. *Practical Relevance of Bipartition Merging*

In theory, the BMA is able to identify rogue taxa for instances that can not be handled by the straight-forward STA (see Section III-B). However, in Section IV-A we observed a substantial agreement between the initial performance curves of the STA and BMA methods. This raises the question how relevant the theoretical advantage of BMA may be on real-world data sets.

To this end, we conducted a more detailed analysis of the drop sets computed by the BMA. To obtain a broader data basis, we also computed the respective drop sets for strict consensus (SC) trees. Fig. 3 depicts the frequency distribution of drop sets according to the number of additional bipartitions that are generated in the consensus by the respective drop set *as well as* the number of taxa that need to be pruned for obtaining these additional bipartition(s). The statistic comprises all drop sets computed across all data sets in our study.

When using MRC and SC in the BMA, drop sets typically consists of a single taxon (MRC: 92.5 %; SC: 89.7 %). In these cases, STA is capable of producing the same result. However, when the drop set contains more than one taxon, STA does not necessarily fail. Thus, for each drop set $D = \{t_1..t_n\}$, we compared the RIC change induced by pruning the full drop set at once (denoted as RIC_{-D}) with the sum of RIC changes, if only one of the taxa in D is pruned (denoted as $\sum \mathrm{RIC}_{-t_i}$). If $\mathrm{RIC}_{-D} > \sum \mathrm{RIC}_{-t_i}$, then this instance can usually not be solved optimally by STA (since analogously to the example in Fig. 1, several taxa *must* be pruned at once). If $\mathrm{RIC}_{-D} < \sum \mathrm{RIC}_{-t_i}$, the drop set is sub-optimal and was only chosen because of an approximation error in the BMA (e.g., a drop set merges three bipartitions into one). Finally, the STA is capable of producing the same result, if $\mathrm{RIC}_{-D} = \sum \mathrm{RIC}_{-t_i}$. For SC, we found that all 12 multi-taxon drop sets (10.3%) require merging, that is, they are not due to approximation errors. For MR, we identified 8 multi-taxon drop sets (1.8%) that can only be optimally solved by the BMA and not by STA. The multi-taxon drop sets occur in data sets with $\geq 1,481$ taxa. On the other hand, we have 15 cases of approximations errors for the MR threshold (3.4%). For these instances, the STA yields more optimal results because it computes RBIC changes explicitly.

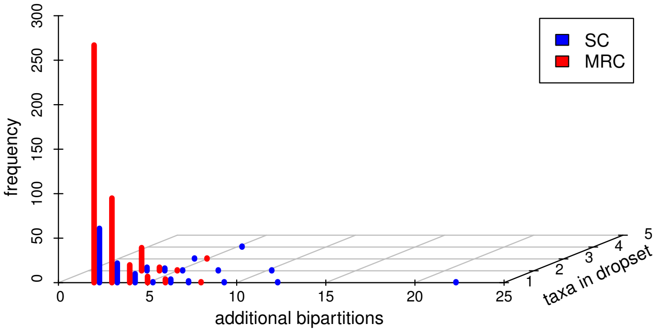


Figure 3: Statistics for all drop sets that are pruned by the BMA when applied to all data sets. On the y-axis, absolute frequencies are depicted for drop sets that add x bipartitions to the consensus tree while pruning z taxa. Drop set frequencies for a MRC threshold are depicted in red, blue bars represent drop set frequencies for a SC.

C. *Necessity of a* de novo *Bootstrap Analysis*

One key issue when pruning taxa from bootstrap trees (and MSAs) is that the consensus tree computed on the pruned bootstrap tree collection may be different from the consensus tree constructed on re-computed bootstrap trees on a pruned MSA (that may also contain fewer alignment sites). In other words, we ask whether it is necessary to re-compute bootstrap trees after rogues have been identified and pruned or if pruning them from the existing tree collection is sufficient.

We address this question by pruning rogues identified by STA from the corresponding MSAs and recomputing bootstrap trees (and subsequently corresponding consensus trees) on the pruned MSAs. We chose not to re-align the sequence data, because initial alignments had been created using various MSA tools and many of them were refined by manual inspection. Thereby, we avoid an alignment method-specific bias in these experiments. With re-alignment it would be specifically hard to disentangle which topological

#initTaxa	$\mathrm{RF_{obs}}$	$\mathrm{p_{pru}}$	$\mathrm{p_{rec}}$	$\mathrm{z_{rec\text{-}pru}}$
24	2	≤ 0.001	≤ 0.001	1.99
44	6	≤ 0.001	≤ 0.001	0.63
52	4	≤ 0.001	≤ 0.001	-1.28
72	5	≤ 0.001	≤ 0.001	-0.14
88	3	0.01	0.06	0.49
125	1	0.54	0.01	1.03
128	3	0.06	0.01	1.29
141	13	≤ 0.001	≤ 0.001	5.52
143	3	0.19	0.02	0.51
148	11	≤ 0.001	≤ 0.001	1.30
150	7	≤ 0.001	≤ 0.001	0.03
218	24	≤ 0.001	≤ 0.001	0.71
316	17	≤ 0.001	≤ 0.001	-1.36
317	11	≤ 0.001	≤ 0.001	-0.39
350	10	≤ 0.001	≤ 0.001	-0.48
354	7	0.01	≤ 0.001	-0.47
404	34	≤ 0.001	≤ 0.001	2.02
424	31	≤ 0.001	≤ 0.001	-0.78
451	33	≤ 0.001	≤ 0.001	0.13
500	22	≤ 0.001	≤ 0.001	0.83
628	36	≤ 0.001	≤ 0.001	0.80
714	51	≤ 0.001	≤ 0.001	1.48
885	44	≤ 0.001	≤ 0.001	0.59

Table II: RF-distances (**$\mathrm{RF_{obs}}$**) between pruned bootstrap trees and *de novo* bootstrap trees computed from pruned MSAs for data sets with up to 885 taxa (**#initTaxa**). P-values (**$\mathrm{p_{pru}}$** and **$\mathrm{p_{rec}}$**) indicate the probability that the RF-distance of two collections of 1,000 trees of $\mathcal{T}_{pru}^{10K}$, resp. $\mathcal{T}_{rec}^{10K}$ (see text), is equal or greater than $\mathrm{RF_{obs}}$. **$\mathrm{z_{rec\text{-}pru}}$** is the z-value of $\mathrm{RF_{obs}}$ in a distribution of RF-distances between consensus trees built from 1,000 trees of $\mathcal{T}_{pru}^{10K}$ and 1,000 trees of $\mathcal{T}_{rec}^{10K}$.

changes in the bootstrap tree collection are due to re-alignment and which are due to re-computation of the trees on a pruned, but otherwise unchanged, MSA.

We denote consensus trees obtained from the original (pruned) bootstrap tree sets as C_{pru} and consensus trees from re-computed bootstrap replicates (*de novo* replicates) on the pruned MSAs as C_{rec}. We computed the symmetric Robinson-Foulds (RF) distance [23] between C_{pru} and C_{rec} (see column $\mathrm{RF_{obs}}$ in Table II). Because of the high computational requirements, we constrained this analysis to data sets with up to 885 taxa. Over all data sets we tested if C_{pru} and C_{rec} are significantly more different from each other with respect to the RF distance, than two consensus trees created under identical conditions.

Therefore, we initially increased the number of bootstrap replicate trees on the original, comprehensive MSAs, to 10,000 and subsequently pruned rogue taxa (the result is denoted as $\mathcal{T}_{pru}^{10K}$). Then, we randomly drew tree set pairs of 1,000 trees each from $\mathcal{T}_{pru}^{10K}$ without replacement and measured the RF-distance between the respective consensus trees. This procedure was repeated 1,000 times to obtain a distribution of 1,000 randomized pair-wise RF-distances. In Table II, $\mathrm{p_{pru}}$ denotes the fraction of RF-distances in the sample distribution being greater or equal to the observed RF-distance $\mathrm{RF_{obs}}$ (see above). Analogously, we also computed 10,000 bootstrap trees on the pruned MSAs (denoted as $\mathcal{T}_{rec}^{10K}$) and created a sample distribution of pairwise RF-distances, once again chosen randomly from the bootstrap tree collections in $\mathcal{T}_{rec}^{10K}$ to obtain a p-value ($\mathrm{p_{rec}}$ in Table II). Finally, we sampled a distribution of 1,000 RF-distances between a consensus tree of 1,000 trees of $\mathcal{T}_{rec}^{10K}$ and a consensus tree of 1,000 trees of $\mathcal{T}_{pru}^{10K}$, to obtain a z-value of the observed $\mathrm{RF_{obs}}$ compared to the sampled distribution (denoted as $\mathrm{z_{rec\text{-}pru}}$ in Table II).

As shown in Table II, the RF-distances between pruned trees and *de novo* computed trees ($\mathrm{RF_{obs}}$) are usually significantly higher than in the null distributions of $\mathrm{p_{pru}}$ and $\mathrm{p_{rec}}$. For almost all data sets, the observed RF-distance lies within the expected variation of RF-distances when the consensus tree of 1,000 pruned bootstrap trees is compared to a consensus tree of 1,000 re-computed bootstrap trees. We conclude that, recomputing bootstrap trees *de novo* on pruned MSAs yields trees that are significantly different from pruned bootstrap trees. Thus, we suggest that bootstrap replicates need to be recomputed after rogue taxa have been pruned.

V. Conclusion and Future Work

We have introduced and made available STA, a novel, simple, and memory-efficient approach for rogue taxon identification. Compared to BMA, it does not only add more bipartitions to the consensus tree, but can also increase support of existing bipartitions by pruning taxa. Furthermore, our algorithm is accurate, in the sense that it only produces sub-optimal results when multi-taxon drop sets are required for increasing support. As shown on a broad and diverse collection of real-world biological MSAs only 1.8 % of the drop sets identified by BMA (using an MR threshold) can not be optimally recovered by STA. On many data sets, we unraveled a substantial potential for improving bipartition support in consensus trees, when a liberal stopping criterion (as implemented in STA) is deployed.

Our study covers a broad spectrum of phylogenetic studies (for further information see [24]) and MSA sizes ranging between 24 and 2,554 taxa. Furthermore, we compared STA and BMA (both integrated in RAxML) with two broadly used taxon stability criteria. TII and LSI perform significantly worse than STA and BMA. Potentially more accurate iterative versions of TII and LSI did not exhibit better performance than their monolithic counterparts. There is a trend for LSI to produce less optimal results than TII with respect to RBIC scores. Given the progressive performance degradation of the TII and LSI on data sets with more taxa, we conclude that there is a tendency for rogue taxa to be unstable. However, rogue taxa are not necessarily the most unstable taxa in the bootstrap tree collection. Therefore, under the optimality criteria used here, node stability methods do not appear to be well-suited for identifying rogue taxa.

Finally, we addressed the question whether it is necessary to conduct a *de novo* bootstrap analysis on pruned MSAs. To this end, we determined RF-distances between consensus trees from pruned bootstrap trees and *de novo* bootstrap trees. Based on a statistical analysis, we find that differences between consensi relying on pruned trees and *de novo* trees on pruned MSAs are significantly larger than expected by chance. Hence, a *de novo* bootstrap inference after rogue identification and removal seems to be mandatory.

We plan to extend STA for optimization of bipartition support in ML/MP trees (drawn from a bootstrap tree collection) and to integrate support for different consensus tree thresholds (i.e., MREC and SC). Beside that, we are working on improving the runtime efficiency of the STA.

Acknowledgment

We are grateful to all biologists, who provided problematic data sets and feedback for this analysis.

References

[1] M. Holder and P. O. Lewis, "Phylogeny estimation: traditional and Bayesian approaches." *Nat Rev Genet*, vol. 4, no. 4, pp. 275–284, Apr. 2003.

[2] J. Felsenstein, "Confidence Limits on Phylogenies: An Approach Using the Bootstrap," *Evolution*, vol. 39, no. 4, pp. pp. 783–791, 1985.

[3] B. Chor and T. Tuller, "Maximum likelihood of evolutionary trees: hardness and approximation." *Bioinformatics*, vol. 21 Suppl 1, pp. i97—-106, Jun. 2005.

[4] W. Day, "The computational complexity of inferring rooted phylogenies by parsimony," *Mathematical Biosciences*, vol. 81, no. 1, pp. 33–42, Sep. 1986.

[5] D. Bryant, "A classification of consensus methods for phylogenetics," in *Bioconsensus: DIMACS Working Group Meetings on Bioconsensus: October 25-26, 2000 and October 2-5, 2001, DIMACS Center*. Amer Mathematical Society, 2003, p. 163.

[6] M. Wilkinson, "Common Cladistic Information and its Consensus Representation: Reduced Adams and Reduced Cladistic Consensus Trees and Profiles," *Systematic Biology*, vol. 43, no. 3, pp. 343–368, 1994.

[7] ——, "More on Reduced Consensus Methods," *Systematic Biology*, vol. 44, no. 3, pp. pp. 435–439, 1995.

[8] ——, "Majority-rule reduced consensus trees and their use in bootstrapping." *Mol Biol Evol*, vol. 13, no. 3, pp. 437–444, 1996.

[9] M. J. Sanderson and H. B. Shaffer, "Troubleshooting molecular phylogenetic analyses," *Annual Review of Ecology and Systematics*, vol. 33, no. 1, pp. 49–72, 2002.

[10] R. C. Thomson and H. B. Shaffer, "Sparse supermatrices for phylogenetic inference: taxonomy, alignment, rogue taxa, and the phylogeny of living turtles." *Syst Biol*, vol. 59, no. 1, pp. 42–58, Jan. 2010.

[11] D. Pisani, A. M. Yates, M. C. Langer, and M. J. Benton, "A genus-level supertree of the Dinosauria," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 269, no. 1494, p. 915, 2002.

[12] C. W. Dunn, A. Hejnol, D. Q. Matus, K. Pang, W. E. Browne, S. A. Smith, E. Seaver, G. W. Rouse, M. Obst, G. D. Edgecombe, M. V. Sø rensen, S. H. D. Haddock, A. Schmidt-Rhaesa, A. Okusu, R. M. b. Kristensen, W. C. Wheeler, M. Q. Martindale, and G. Giribet, "Broad phylogenomic sampling improves resolution of the animal tree of life." *Nature*, vol. 452, no. 7188, pp. 745–749, Apr. 2008.

[13] E. A. Sperling, K. J. Peterson, and D. Pisani, "Phylogenetic-signal dissection of nuclear housekeeping genes supports the paraphyly of sponges and the monophyly of Eumetazoa." *Mol Biol Evol*, vol. 26, no. 10, pp. 2261–2274, 2009.

[14] R. C. Thomson and H. B. Shaffer, "Rapid progress on the vertebrate tree of life." *BMC Biol*, vol. 8, p. 19, 2010.

[15] Thorley and Wilkinson, "Testing the phylogenetic stability of early tetrapods," *J Theor Biol*, vol. 200, no. 3, pp. 343–344, 1999.

[16] W. P. Maddison and D. R. Maddison, "Mesquite: a modular system for evolutionary analysis," 2010.

[17] M. Wilkinson, "Identifying stable reference taxa for phylogenetic nomenclature," *Zoologica Scripta*, vol. 35, no. 1, pp. 109–112, 2006.

[18] N. Pattengale, A. Aberer, K. Swenson, A. Stamatakis, and B. Moret, "Uncovering Hidden Phylogenetic Consensus in Large Datasets." *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, vol. X, no. X, pp. 1–11, Feb. 2011.

[19] A. J. Aberer, N. D. Pattengale, and A. Stamatakis, "Parallelized phylogenetic post-analysis on multi-core architectures," *Journal of Computational Science*, vol. 1, no. 2, pp. 107–114, 2010.

[20] N. D. Pattengale, M. Alipour, O. R. P. Bininda-Emonds, B. M. E. Moret, and A. Stamatakis, "How many bootstrap replicates are necessary?" *J Comput Biol*, vol. 17, no. 3, pp. 337–354, 2010.

[21] A. Stamatakis, P. Hoover, and J. Rougemont, "A rapid bootstrap algorithm for the RAxML web servers," *Systematic Biology*, vol. 57, no. 5, p. 758, 2008.

[22] A. Stamatakis, "Phylogenetic models of rate heterogeneity: a high performance computing perspective," in *Proceedings 20th IEEE International Parallel & Distributed Processing Symposium*. IEEE, 2006, p. 278.

[23] D. F. Robinson and L. R. Foulds, "Comparison of phylogenetic trees," *Math. Biosci.*, vol. 53, pp. 131–147, 1981.

[24] A. J. Aberer, "Advanced Methods for Phylogenetic Post-Analysis," Master's thesis, TU/LMU Munich, 2011.