Molecular Phylogenetics and Evolution 105 (2016) 114–125

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Split-specific bootstrap measures for quantifying phylogenetic stability and the influence of taxon selection

Huai-Chun Wang [a,b,c,*], Edward Susko [a,c], Andrew J. Roger [b,c,d]

[a] Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia B3H 4R2, Canada
[b] Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia B3H 4R2, Canada
[c] Centre for Comparative Genomics and Evolutionary Bioinformatics, Dalhousie University, Canada
[d] Canadian Institute for Advanced Research, Program in Integrated Microbial Biodiversity, Canada

ARTICLE INFO

Article history:
Received 19 July 2015
Revised 12 August 2016
Accepted 23 August 2016
Available online 24 August 2016

Keywords:
Phylogenetic stability
Bootstrap support
Split
Taxon influence
Taxon sampling

ABSTRACT

Assessing the robustness of an inferred phylogeny is an important element of phylogenetics. This is typically done with measures of stabilities at the internal branches and the variation of the positions of the leaf nodes. The bootstrap support for branches in maximum parsimony, distance and maximum likelihood estimation, or posterior probabilities in Bayesian inference, measure the uncertainty about a branch due to the sampling of the sites from genes or sampling genes from genomes. However, these measures do not reveal how taxon sampling affects branch support and the effects of taxon sampling on the estimated phylogeny. An internal branch in a phylogenetic tree can be viewed as a split that separates the taxa into two nonempty complementary subsets. We develop several split-specific measures of stability determined from bootstrap support for quartets. These include BPtaxon_split (average bootstrap percentage [BP] for all quartets involving a taxon within a split), BPsplit (BPtaxon_split averaged over taxa), BPtaxon (BPtaxon_split averaged over splits) and RBIC-taxon (average BP over all splits after removing a taxon). We also develop a pruned-tree distance metric. Application of our measures to empirical and simulated data illustrate that existing measures of overall stability can fail to detect taxa that are the primary source of a split-specific instability. Moreover, we show that the use of many reduced sets of quartets is important in being able to detect the influence of joint sets of taxa rather than individual taxa. These new measures are valuable diagnostic tools to guide taxon sampling in phylogenetic experimental design.

1. Introduction

Phylogeneticists are increasingly utilizing large numbers of orthologous genes from many taxa to infer the evolutionary relationships among organisms. These phylogenomic approaches, based either on estimation from concatenated alignment of hundreds of genes as a single dataset, or supertree methods applied to hundreds of estimated individual gene trees, can drastically reduce stochastic errors associated with small datasets used in traditional phylogenetic studies and have led to substantial advances in our knowledge of the tree of life (Delsuc et al., 2005). However, even with large amounts of data there can be problems with apparent lack of support in estimated trees. For example, in analyses of super-matrices of greater than 100 genes addressing global relationships among eukaryotes, the relationships among some groups, such as Archaeplastida, Haptophytes and Cryptophytes remain poorly supported or are differently resolved depending on the dataset used (Burki et al., 2012; Yabuki et al., 2014).

One possible source of this apparent lack of resolution or conflict in phylogenetic datasets is unstable taxa (also called rogue taxa or 'rogues' by Wilkinson, 1996 or 'wildcard taxa' and 'floating taxa' as in Goloboff and Szumik, 2015). Unstable taxa may contain missing data (Wilkinson, 1995), have an elevated substitution rate causing homoplasy (Sanderson and Shaffer, 2002), or manifest compositional heterogeneity (Foster, 2004), selection (Thomas, 2007) and different evolutionary histories for different data partitions that are not adequately addressed by phylogenetic models. They tend to 'bounce around' to different positions in the estimated trees of bootstrap replicates, providing low support for some branches and/or disrupting many relationships that are otherwise well resolved. The removal of rogue taxa has thus become a common practice in phylogenomic studies and consensus summaries of the trees are obtained to maximize the overall

[*] Corresponding author at: Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia B3H 4R2, Canada.
E-mail address: huaichun.wang@dal.ca (H.-C. Wang).

http://dx.doi.org/10.1016/j.ympev.2016.08.017
1055-7903/© 2016 Elsevier Inc. All rights reserved.

stability of the phylogeny (Wilkinson, 1994, 1996; Goloboff and Farris, 2001; Swenson et al., 2011). In some cases the whole analysis will have to be re-conducted with the rogue taxa removed (Thomson and Shaffer, 2009; Aberer and Stamatakis, 2011; see also Goloboff and Szumik, 2015 for a discussion of pruning trees or matrices). Several methods for rogue detection have been developed including: reduced consensus trees (Wilkinson, 1994, 1996), the leaf stability index (LSI; Thorley and Wilkinson, 1999; Wilkinson, 2006), various consensus (and other) methods implemented in Tree analysis using New Technology (TNT; Goloboff et al., 2008), IterPCR (Pol and Escapa, 2009), taxonomic instability index (TII; Maddison and Maddison, 2010), relative bipartition information content (RBIC; Aberer and Stamatakis, 2011) and consensus network (Holland and Moulton, 2003; Huson and Bryant, 2006). Goloboff and Szumik (2015) discussed much of the early work on the identification of unstable taxa and demonstrated the algorithmic efficiency of TNT in implementing the LSI measure compared with Phyutility (Smith and Dunn, 2008).

The stability of an individual taxon can be a useful indicator of its overall influence on the estimated phylogenetic tree. However, some of the often used taxon stability measures (e.g., LSI and TII) cannot detect groups of jointly influential taxa. Moreover, such overall stability measures may not detect taxa that are influential for a specific split (an internal edge) in a phylogeny. The stability of an internal edge is most often measured with nonparametric bootstrapping (Felsenstein, 1985), but jackknifing (Siddall, 1995; Farris et al., 1996) or Bayesian posterior probabilities (Huelsenbeck et al., 2000) are also used. The bootstrap support (BP) for an edge corresponds to the percentage of bootstrap trees that contain such an edge (bipartition or split of the taxa). It does not provide any information about the other bipartitions that conflict with the current edge or the relative support of each taxon for the edge. To address this problem, other measures of internal edge stability have recently been developed (Salichos and Rokas, 2013; Salichos et al., 2014; Mariadassou et al., 2012; Sheikh et al., 2013).

The minimal subset of taxa whose evolutionary relationship defines an internal edge in a larger unrooted tree is four. By considering how bootstrap support varies across all four-taxon trees that uniquely define the edge of interest, we are able to detect groups of taxa whose joint effects are localized to splits. We propose novel, robust measures of phylogenetic stability that can be used to identify taxa that are responsible for high or low support for branches. We use several simulated datasets plus one empirical dataset to illustrate the utility, properties and importance of these split-specific measures and compare them to several existing measures of phylogenetic stability.

2. Materials and methods

2.1. Influence measures from quartets

The measures that we consider here detect taxa that have a large influence on statistical support for an edge on a tree. Measures of overall influence on a tree are obtained by averaging edge-specific measures of influence. A straightforward way of investigating whether a taxon or taxa have a large influence on support for an edge is to examine support with and without the taxon or taxa. Individual taxa need not be influential on their own, however. Groups of closely related taxa can be expected to be jointly influential. To be able to find groups of influential taxa, one must delete enough taxa for the importance of their absence to show in changes of support. The greatest reduction in taxa that still provides information about an edge of interest leaves just four taxa. Not all subsets of four taxa will provide information specifically about the edge, however. To provide information, their four-taxon tree must contain the edge. Given a tree of the full set of taxa and an edge of interest, the procedure here determines all sets of four taxa quartets whose subtree has a middle edge that coincides with the edge of interest.

To illustrate the procedure, it is convenient to relate the edges in a tree with splits (taxa bipartition). For instance, edge 8 in Fig. 1A has corresponding split representation ABCFG|ED. Every internal edge in a bifurcating phylogenetic tree has four adjacent edges. Let the split for an internal edge of interest be S|S′ and let the splits for its adjacent edges be $\mathrm{S_j|S'_j}$, $j = 1, \ldots, 4$. Exactly one of $\mathrm{S_j}$ or $\mathrm{S'_j}$ is contained in either S or S′. Because labeling is arbitrary, we denote $\mathrm{S_j}$ as the set contained in S or S′. For instance, in Fig. 1A, the edges adjacent to 8 are 7, 10 and the two terminal edges leading to E and D. The corresponding $\mathrm{S_j}$ are {G, F}, {A, B, C}, {E} and {D}. We refer to $\mathrm{S_1}, \ldots, \mathrm{S_4}$ as the four quartet subgroups for the split. The sets of four taxa of interest are obtained by selecting taxa A1, . . . , A4 from each of S1, . . . , S4. The tree for A1, . . . , A4 can equivalently be expressed as a quartet A1A2|A3A4. We refer to the collection of quartets over all choices of A1, . . . , A4 as the quartets for the edge. For example edge 9 in Fig. 1A gives rise to the following quartets: AB|CD, AB|CE, AB|CF and AB|CG. However, not any group of four taxa is a quartet in our definition. AB|FG, for instance, is not a quartet for edge 9 only but corresponds to the sum of edges 7, 9 and 10. Methods that include AB|FG among quartets would check for influence on any one of splits 7, 9 or 10 rather than just split 9 and thus are not split-specific.

We consider several measures of influence based on a list of quartets sampled from a given tree of interest. The frequency of the quartets are derived from a set of bootstrap trees:

(1) BPtaxon_split: For a given edge and taxon of interest, average BP of all quartets for that edge that involve the taxon, which is a measure of the relative support of individual taxa for the split.
(2) BPsplit: For a given edge of interest, BPtaxon_split averaged over taxa, which is an overall measure of the internal split stability.
(3) BPtaxon: For a given taxon of interest, BPtaxon_split averaged over splits, which is an overall measure of stability of the taxon. This measure is similar to the LSI (Thorley and Wilkinson, 1999) except that BPtaxon only considers split-specific quartets.

It can be shown that the following inequality holds:

$$\mathrm{BP} \leqslant \mathrm{BPsplit} \tag{1}$$

To see this, consider a split $s = S|S'$ in a reference tree of interest, and an arbitrary quartet $a_1\,a_2|a_3\,a_4$ that corresponds to that split where a_1, a_2 in S and a_3, a_4 in S'. For the ith bootstrap tree, τ_i, if s is in τ_i, then there is an edge where all of the taxa in S are separated from all of the taxa in S'. Since a_1, a_2 in S and a_3, a_4 in S', the quartet $a_1\,a_2|a_3\,a_4$ is displayed in τ_i. In other words, the indicator function $I\{a_1\,a_2|a_3\,a_4 \text{ in } \tau_i\} = 1$, whenever $I\{s \text{ in } \tau_i\} = 1$. Since $I\{a_1\,a_2|a_3\,a_4 \text{ in } \tau_i\}$ is either 0 or 1, whenever $I\ \{s \text{ in } \tau_i\} = 0$, $I\{a_1\,a_2|a_3\,a_4 \text{ in } \tau_i\} \geqslant I\{s \text{ in } \tau_i\}$. Summarizing the two cases we have,

$$I\{s \text{ in } \tau_i\} \leqslant I\{a_1\,a_2|a_3\,a_4 \text{ in } \tau_i\}$$

Summing over i and dividing by B (the number of bootstrap trees) gives

$$BP = \frac{1}{B}\sum_i I\{s \in \tau_i\} \leqslant \frac{1}{B}\sum_i I\{a_1\,a_2|a_3\,a_4 \in \tau_i\}$$

Averaging over all set of quartets for a given taxon gives that BP ⩽ BPtaxon_split and further averaging over all taxa gives that BP ⩽ BPsplit.

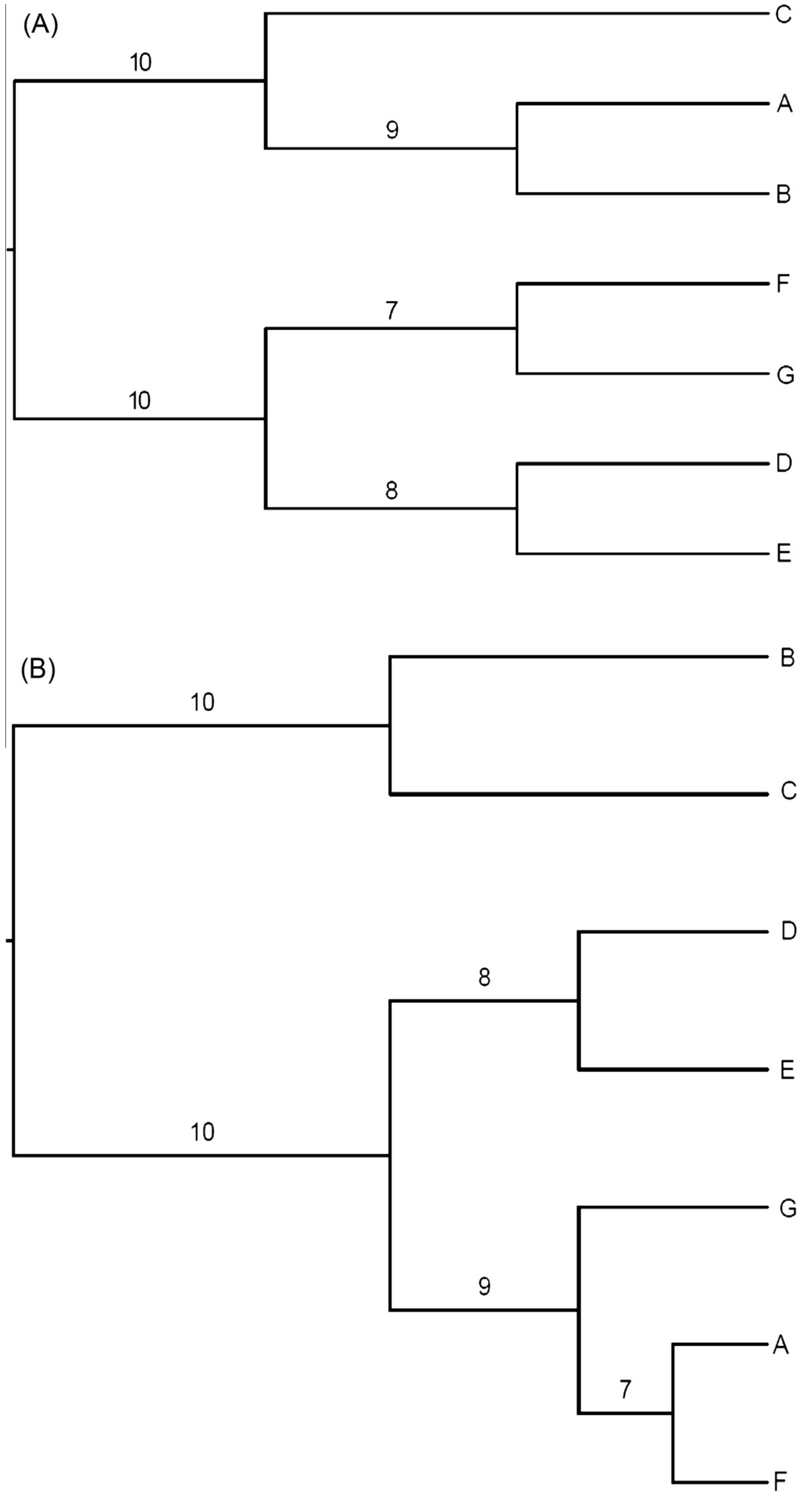

Fig. 1. (A) and (B) Two 7-taxa trees with internal edges (splits) labeled with numbers. The external edges (leading to a taxon) are labeled from 0 to $T-1$ (not shown); the internal edges are numbered from T to $2T-4$ ($T = 7$ in this case).

For instance, in Fig. 1A, BP for split 8 (ED|ABCFG) is the frequency of the bootstrap trees that contain this bipartition, which is less than or equal to the corresponding BPsplit, being the average frequency of the six types of quartets arising from the split: GA|ED, GB|ED, GC|ED, FA|ED, FB|ED and FC|ED. The equality holds only when the average frequencies of each kind of the quartets are all equal and equal to the BP, which means all taxa have equal contribution to the frequency of each kind of the quartets (BPtaxon_split = BP). The inequality between BP and BPsplit is consistent with the decay theorem (Wilkinson et al., 2000) that proves that the support for a phylogenetic hypothesis (such as a split) is no greater than the lowest support for any of the less inclusive hypotheses (such as quartets) that it entails.

2.2. The relative bipartition information content (RBIC)

Removing taxa from a set of bootstrap trees will likely merge some bipartitions with each other to form new bipartitions or let some of them disappear and thus alter the number and supports of bipartitions included in the consensus tree or a tree of interest. For instance, consider the trees in Fig. 1 as two bootstrap trees (Fig. 1A and B). The split labeled 10 is different for the two trees, but once taxon A is removed, the new split in the resulting two trees that corresponds to split 10 is the same. The RBIC for a consensus tree is the sum of the relative frequencies of all splits in the bootstrap trees and divided by $T-3$ where T is the number of the taxa in the initial taxon set (Aberer and Stamatakis, 2011). Aberer and colleagues developed several methods, implemented in their software (RogueNaRok) to maximize the RBIC score for a consensus tree when taxa are pruned to detect rogue taxa (Aberer et al., 2013). Because RogueNaRok does not output RBIC values which depend on consensus trees, we used, in its place, the average bootstrap frequency over all splits still present in the tree of interest, after the taxon is removed. We call it RBIC-taxon, which differs from the RBIC in RogueNaRok by a scalar multiple in most cases.

A complication for the RBIC measure is that the nature of splits changes when taxa are removed. Two cases need to be considered: in the first case, the taxon deleted is part of a cherry (two sister taxa at the tip of a branch) and removal of either of the two taxa will make their parent split a terminal (trivial) split. RBIC for such combinations of splits and taxa are ignored. The second case occurs when the taxon deleted joins the tree at an internal edge; for instance taxon C in Fig. 1A. After deleting this taxon, the node it connects to disappears and the two neighboring branches merge into a single branch. There are two ways of dealing with this in RBIC calculations. First, one can simply ignore the RBIC for this specific taxon – split combination on the grounds that inclusion of the resulting 'single merged branch' in RBIC calculations is for a sum of branches, rather than a single branch. Another way is not to ignore this case but instead to avoid double counting its contribution to RBIC-taxon. In this case, the branch should be counted for one but not both of the branches next to the removed taxon. For example, for the tree in Fig. 1A RBIC for split 9 or 10 might be considered (it does not matter which) when C is deleted but not both. In this work we used the second approach, which is the same as in Aberer and Stamatakis (2011), to calculate the RBIC for these cases.

2.3. Tree distances between a reference tree and bootstrap trees

Another metric for the stability of a taxon on a phylogeny is the average distance between a reference tree (e.g., the ML tree) with the taxon pruned and the bootstrap trees with the same taxon pruned (Eq. (2)). The symmetric Robinson-Foulds (RF) distance between a pair of trees (Robinson and Foulds, 1981) is used, which is the sum of the number of splits that are unique to either one of the two compared trees. We refer to the resulting measure of taxon stability as a pruned-tree RF distance metric:

$$d(t)=\frac{\sum_{i=1}^{B}\Delta\left(\tau_i^{(-t)},\tau_0^{(-t)}\right)}{B} \tag{2}$$

where B is the number of bootstrap trees, $\Delta\left(\tau_i^{(-t)},\tau_0^{(-t)}\right)$ is the RF distance between a bootstrap tree τ_i and reference tree τ_0 after removal of taxon t from both trees.

The rationale for this measure is that if the relative position of a taxon is stable, the bootstrap trees should be similar to the ML tree, whereas rogue taxa will appear in different positions of the bootstrap trees and thus increase the RF distance between the bootstrap trees and the ML tree. Removing the rogue taxa will reduce this effect and so reduce the average RF distance, while removing stable taxa cannot result in reduction in the average RF distance.

All the above five measures were calculated for the test datasets and compared with several existing measures for phylogenetic stability. For taxon stabilities, the latter include the LSI, TII and RogueNaRok, all of which were calculated using the RogueNaRok software package (Aberer et al., 2013). The LSI, calculated as the difference in bootstrap frequencies between the most frequent quartets and the second most abundant quartets involving a taxon (Thorley and Wilkinson, 1999), was used through this study. For internal edge stabilities, two recently proposed measures, internode certainty (IC) and internode certainty all (ICA) (Salichos and Rokas, 2013; Salichos et al., 2014), were calculated with RAxML 8.0 (Stamatakis, 2006), in addition to the commonly used bootstrap percentage (BP) score. The IC and ICA for a split (or internode) are defined as 1 minus the Shannon entropy of the abundancies of a split of interest and its most prevalent conflicting split and 1 minus the entropy of the abundancies of the split of interest and all of its conflicting splits in a given set of trees, respectively (Salichos et al., 2014). IC and ICA scores at or close to 1 indicate the absence of conflict for the bipartition defined by the internal edge and the BP for the edge is consequently at or close to 100%; whereas IC and ICA at or close to 0 indicate equal support for the conflicting splits and hence maximum conflict at the edge, and the BP for the edge is around 50% when there are only two conflicting splits. IC is determined from the relative bootstrap support of the split: the proportion of times it arises in bootstrap trees among trees that contain either it or the most frequently occurring split that is incompatible with it. We will call the relative bootstrap support as *relative* BP to distinguish it from the standard BP score for a split. The relative BP is similar to the "GC (Group present/Contradicted)" consensus (Goloboff et al., 2003) and the LSI (Thorley and Wilkinson, 1999). As long as the relative BP for the split of interest is larger than 50%, IC is uniquely determined by it. An explicit formula for calculating the relative BP from IC is not available but it can be determined numerically from IC. Since the two most frequently occurring splits can, by definition, be expected to occur frequently, relative BP tends to correlate strongly with BP and is the same as BP whenever the bootstrap trees only contained the split of interest or its most prevalent conflicting split. A large difference between the relative BP and BP often suggests that there are several conflicting splits for the split of interest in the bootstrap trees.

2.4. Test data

The first set of data were generated by simulating sequence evolution over a 14-taxon tree (Fig. 2), two external branches of which were very long but with short internal branches chosen so that a long-branch attraction (LBA) artifact would be induced. The program seq-gen (Rambaut and Grassly, 1997) was used to simulate one hundred replicated datasets of 20,000 nucleotide sites each under a nucleotide JC69 model (Jukes and Cantor, 1969) and a continuous Gamma distribution of site-wise rates of evolution ($\alpha = 1.0$). An ML tree and 1000 bootstrap trees were estimated with RAxML under a nucleotide GTR + Γ model for each bootstrap replicate, which we call the SimuLBA data.

The second dataset we considered consists of 22 birds and 2 reptile mitochondrial genome sequences (Phillips et al., 2010), which will be called the ratites data. The bird species include ten ratites (a group of large flightless birds such as ostrich and emu) and 12 other birds. The two reptiles are alligator and caiman used as the outgroup. The DNA sequence alignment (14,190 nucleotide sites) together with the information about the partition of the sites into five groups and 1000 bootstrap trees, estimated under a GTRCAT model with RAxML (Stamatakis, 2006), were downloaded from the data repository website indicated in Aberer and Stamatakis (2011). A ML tree, not available from the previous

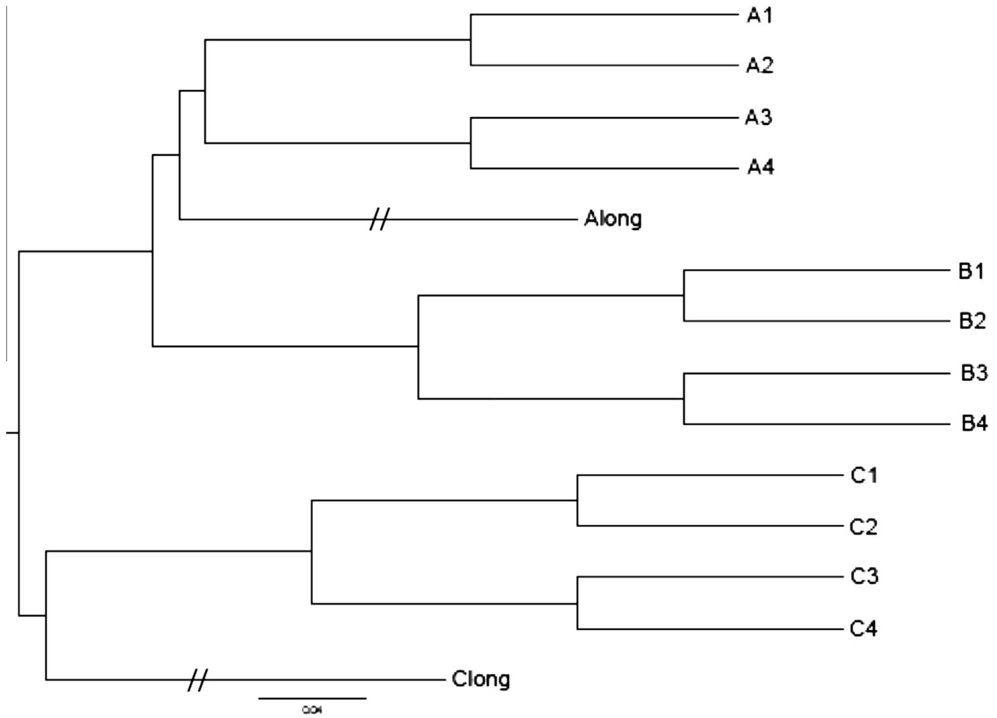

Fig. 2. A 14-taxon tree used to simulate DNA sequence data under a JC69 + Γ model. The edges leading to Along and Clong, shortened here for clarity of the tree, have branch length of 3.0 each.

source, was estimated under the same model, including site partitions, using RAxML.

3. Results

3.1. The SimuLBA data

Of the one hundred replicated nucleotide datasets simulated under a 14-taxa tree that contains two non-sister long branches (Along and Clong) (Fig. 2), the ML trees estimated under a GTR + Γ model had the same topology as the generating tree in 34 datasets. For 35 other datasets the estimated ML trees were LBA-biased where Along and Clong formed a split separate from the other taxa. For the other 31 datasets the ML trees had slightly different topologies from the generating tree. Specifically, the position of Along was incorrect: Along formed a split with C1–C4 in 4 datasets; it formed a split with A1/A2, A3/A4 and B1–B4, separately in 9 datasets each. The following two sections examine in detail two splits of interest whose presence determines whether the estimated tree is LBA biased or not.

3.1.1. The split of C1–C4 & Clong from the other taxa: the trees are not LBA-biased

This split exists in the ML trees estimated from 65 datasets (Fig. 3A). For these datasets, when BP was small, BPsplit tended to be large. In only 11 cases was the difference between BP and BPsplit less than 1%. Thus BPsplit more strongly supported the correct split than BP. Cases where the four quartet subgroups for the split coincided with those in the true tree (marked with a 1 in the figure) tended to give rise to a larger gap between BP and BPsplit. Fig. 4A shows one of the trees estimated from dataset 2 with split number, BP, BPsplit and IC scores labeled for the three splits that are not 100% resolved, including split 19 separating C1–C4 & Clong from the other taxa. The BP score for split 19 was low (52%). BPsplit, however, was much higher at 87%, lending much greater support for this correct branch. The IC and ICA scores for this split are both 0.01, which means this split and its conflicting split (that supports a LBA tree) have nearly equal bootstrap support. The relative BP for the branch, deduced from the IC (see Section 2), is 56%, which is slightly greater than the BP of 52%, suggesting that there are several splits that conflict with it. Here the relative BP and hence the IC and ICA, like BP, failed to provide strong support for this branch, whereas BPsplit correctly gave a much stronger support.

Similarly, for the tree of dataset 10, BPsplit, BP, IC, ICA for the C1–C4 & Clong split are 64%, 54%, 0.004 and 0.004, respectively. As a rough rule, IC and ICA scores close to 0 often correspond to relative BPs approximately equal to 50%. Therefore, BPsplit also gave stronger support than BP and IC for this correct branch.

3.1.2. The split of Along & Clong from the other taxa: the trees are LBA-biased

This split manifesting the LBA artifact is present in 35 ML trees (Fig. 5A). BPsplit equals BP in all of these cases. One of the trees (from dataset 7) is shown in Fig. 4B and the split of interest is labeled as 17. BP and BPsplit for this split are 66% and the IC and ICA values are 0.078 and 0.07 respectively. The relative BP corresponding to this IC is also 66% which is the same as the BP suggesting among all of the bootstrap trees there is only one split that conflicted with split 17.

In Section 2 we have shown mathematically that BPsplit will not be smaller than BP. What is interesting here is that BPsplit is much higher than BP for the correct C1–C4 & Clong split in many of the cases, but BPsplit is not higher than BP for the wrong Along & Clong split in all cases.

3.1.3. The other estimated splits

We looked at the other splits that have BP less than 100%. Two of the correct splits are (A1–A4 & Along) which exists in 52 ML trees and (A1–A4) present in 82 ML trees. Fig. 3B and C shows the scatter plots of BPsplit vs. BP for the two splits respectively.

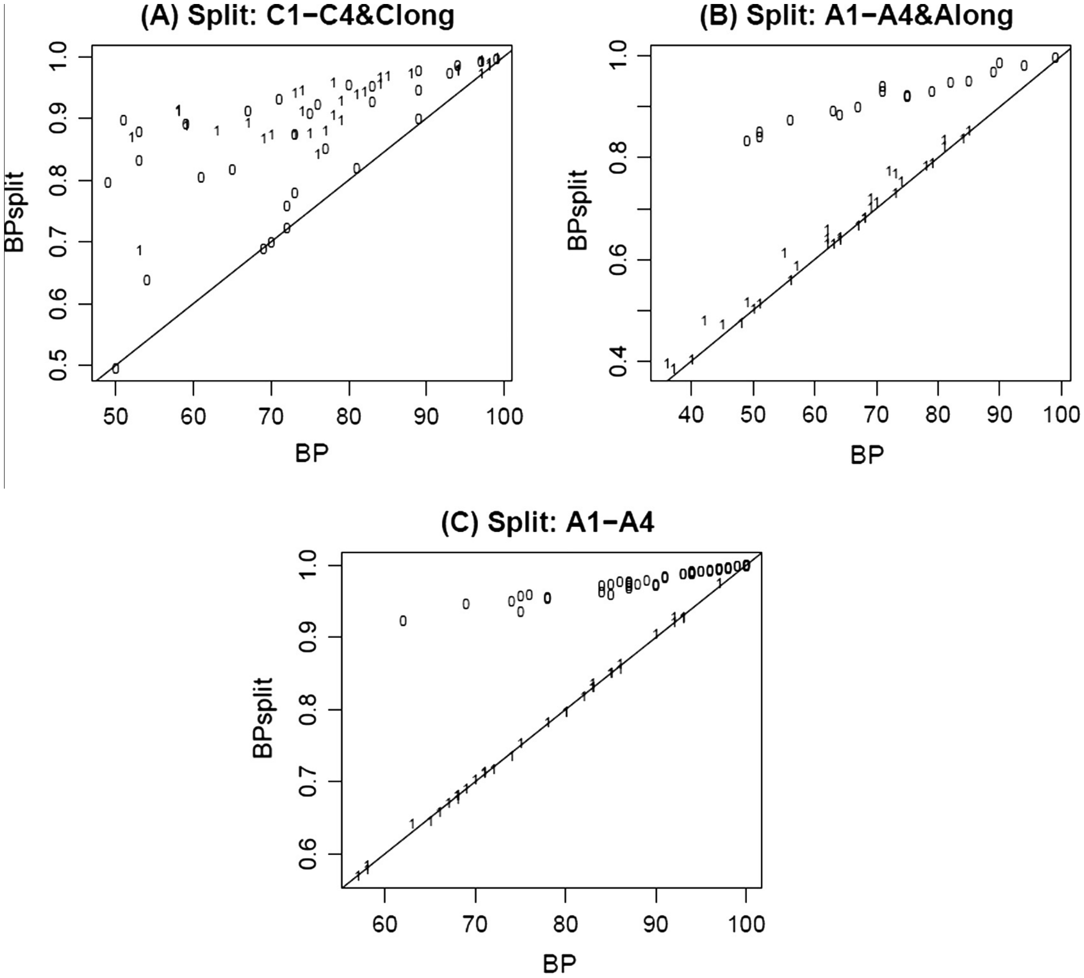

Fig. 3. Scatter plots of BPsplit against BP for the three correctly estimated splits that came up in ML trees with BP < 100%. Labels indicate whether the ML tree had four quartet subgroups for the split that coincided with the true tree (labeled 1) or not.

For a substantial number of datasets that had these splits in the ML tree, BPsplit was much larger than BP, correctly suggesting stronger support for these correct splits. For each such dataset, the ML tree had four quartet subgroups for the split that did not match those of the true tree. Being able to average over quartets that did not include Along was important in detecting that split was more likely to be present than suggested by BP. Trees where the four quartet subgroups differed from the true tree allowed this as Along, by itself, did not constitute a subgroup, which would have caused it to be present in every quartet. Moreover, when we examined the taxa that best agree with these two splits, using the BPtaxon_split measure, we found that, for the cases where BPsplit were much greater than BP, quartets containing Along or Clong always had the least support of all the 14 taxa. For those splits where BPsplit and BP were approximately the same, there was no difference in BPtaxon_split among the taxa.

Among the incorrect splits, split (A1–A4 & B1–B4) exists in 38 datasets; split (A1, A2, Along), split (A3, A4, Along) and split (Along & B1–B4) are present in 9 datasets in each case. Fig. 5B–E shows the scatter plot of BPsplit vs. BP for the 4 splits respectively. Except for split (A1–A4 & B1–B4) where BPsplit is higher than BP in most cases, the two quantities are similar for the other three splits. A fifth wrong split (Along, B1–B4, Clong) is present in one dataset and both BP and BPsplit are 59%.

These results again show that: (1) BPsplit is higher than BP in the majority of the cases for the correct splits and (2) BPsplit is not much different from BP for the wrong splits in most of the datasets. There are some exceptions to the second point, especially for the wrong split (A1–A4 & B1–B4) where BPsplit was much higher than BP, apparently reinforcing support for an incorrect grouping. However, the A1–A4 & B1–B4 split is special in that Along is attracted by Clong to form a wrong split of Along and Clong leaving a suboptimal split of A1–A4 & B1–B4. Indeed among the 38 ML trees that have this split 34 trees also contain the LBA split (Along & Clong) and it is those trees (labeled as ‘2’ in Fig. 5B) that have higher BPsplit than BP for the A1–A4 & B1–B4 split. Furthermore, we examined the ML tree for dataset 7 that has this split (see Fig. 4B). Its IC (and ICA) score is 0.233 (and 0.23), which gives a relative BP for the split of interest of about 78%; and this support level is much closer to BPsplit (82%) than to BP (66%). Taken together, the simulation results demonstrate that in the current simulation scenario where the long branches bounce around the trees, BPsplit is often much larger than BP for the correctly estimated edges, but it is not bigger than BP for many of the incorrect edges, suggesting BPsplit provides a more robust measure of edge supports than the traditional BP measure.

3.1.4. Taxon stability

For the one hundred simulated datasets we applied the five measures of taxon stability including the three that are proposed here (BPtaxon, RBIC-taxon and the pruned-tree RF distance metric) and two existing ones (LSI and TII). Table 1 shows the least and the second least stable taxa according to the methods. All methods detected Along was the most unstable taxon in the majority of

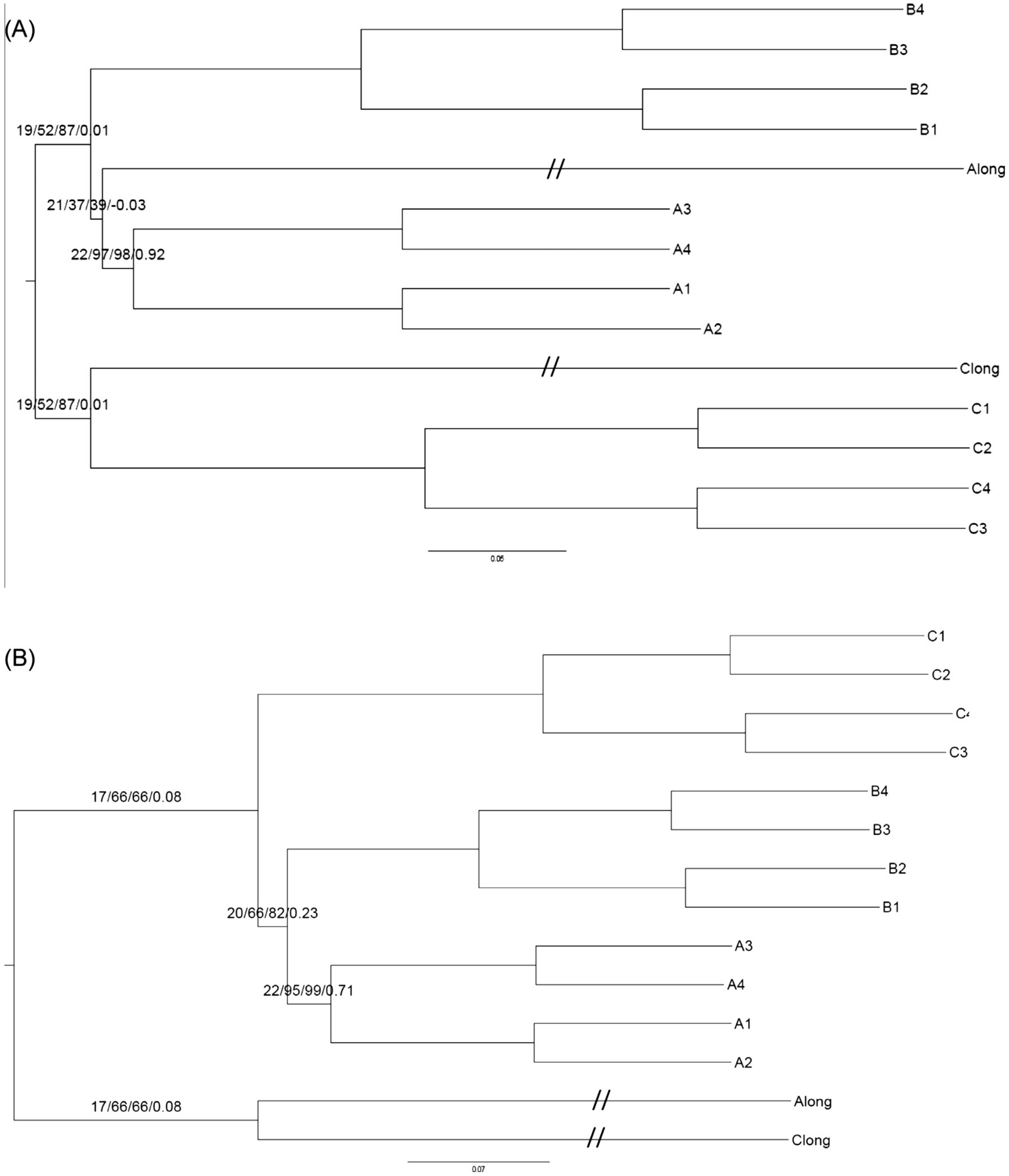

Fig. 4. (A) Correctly estimated ML tree from SimuLBA dataset 2. (B) A ML tree with a long branch attraction bias estimated from SimuLBA dataset 7. For both trees, splits with BP less than 100% are labeled, on which the first number is the split number, followed by BP, BPsplit and IC. For the unlabeled splits both BP and BPsplit are 100% and IC = 1. Both Along and Clong have been shortened by 10-fold in the figures.

the cases followed by Clong. Indeed among the 100 datasets, the position of Along is most variable. In 52 datasets Along forms a correct split with A1–A4; in 35 datasets it forms a wrong LBA split with Clong; and in the rest 13 datasets Along groups with B1–B4 or C1–C4. The RBIC algorithm in the RogueNaRok program detected Along to be rogue taxa in 30 datasets and Clong to be rogue in 10 other datasets.

3.2. The ratites data

3.2.1. Internal split stability

The ML tree re-estimated under a GTRCAT + site partition model with RAxML is the same as published by Phillips et al. (2010). In particular the three tinamou birds are nested within the ratites, although the BP for the tinamous-moa split (split 39 in Fig. 6) is only 60% in the ML tree. The BPsplit for this branch is much higher at 91% and is much closer to the high BP and Bayesian posterior probability (BPP) scores (being 99% and 1.0 respectively) obtained by Phillips et al. (2010) under different phylogenetic models. Similarly split 37 and split 40 both have BP of 26% and their BPsplit are 57% and 36% respectively. These higher BPsplit values are very close to the BP scores for the two branches (56% and 36% respectively) in Phillips et al. (2010) and their BPP scores are even higher. The five edges with the smallest BPsplit also had the smallest IC scores (Fig. 6). In particular, the two edges (37 and 40) with the smallest BPsplit had negative IC scores, indicating that conflicting splits had higher bootstrap percentages than the splits present in the ML tree.

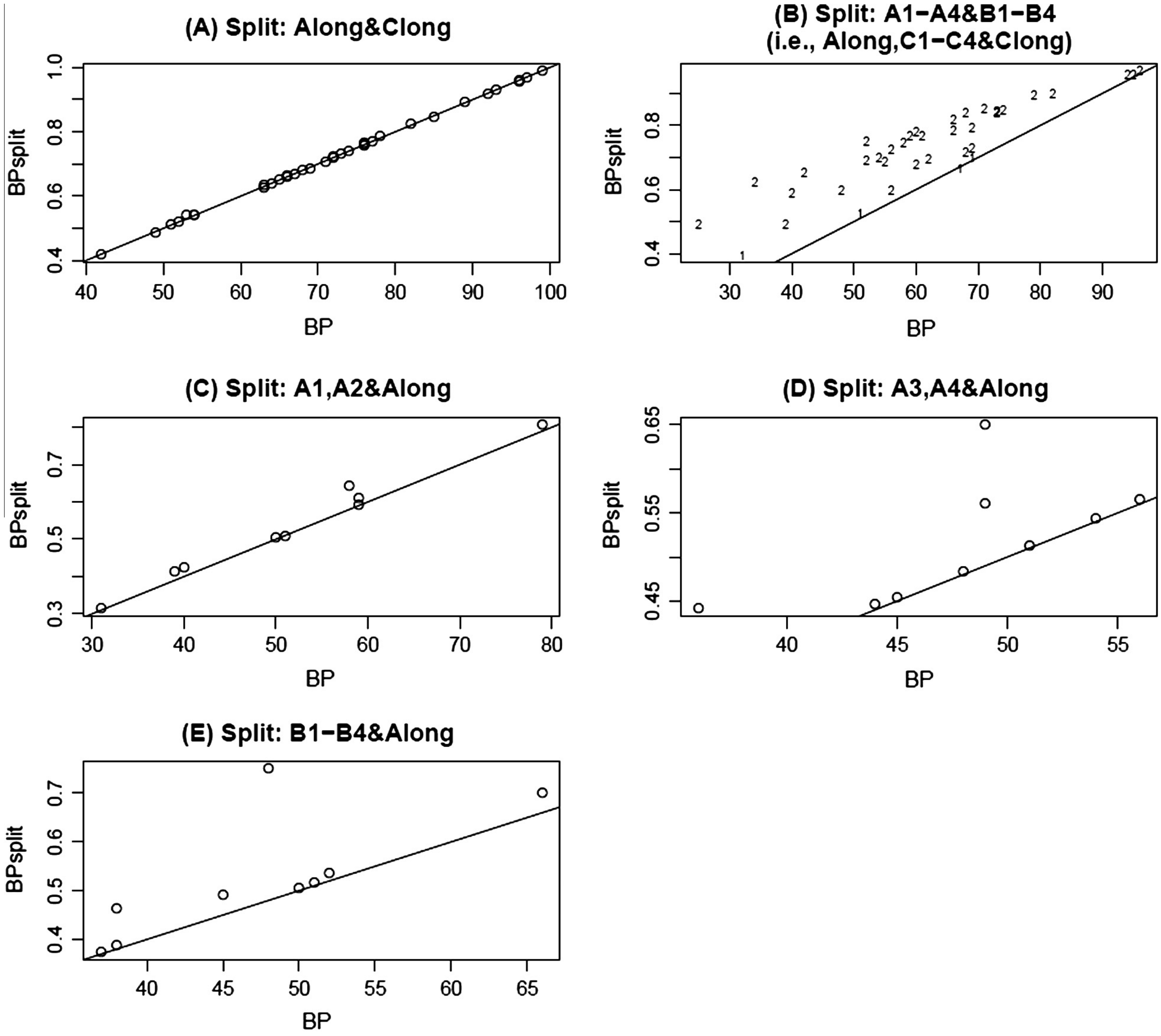

Fig. 5. Scatter plots of BPsplit vs. BP for five incorrectly estimated splits. Circles in (A), (C), (D) and (E) are the datasets. The numbers in (B) represent the ML trees from the datasets contain a C1–C4 & Clong split (labeled 1) or an Along & Clong split (labeled 2).

Table 1
Number of the least and second least stable taxa in the 100 simulated datasets under five taxon stability measures.

Taxon stability measure	Least stable taxon		Second least stable taxon	
	Along	Clong	Along	Clong
BPtaxon	60	38	27	37
LSI	84	16	16	52
TII	82	18	18	71
RBIC-taxon	86	14	14	85
Pruned-tree RF distance	86	14	14	85

To determine which taxa were responsible for high or low support values for a particular split, we calculated BPtaxon_split for each taxon. Of the 21 splits in the 24-taxon tree, five had BPsplit values less than 0.95, among which split 30 had a constant BPtaxon_split score across taxa; the other four splits (34, 37, 39 and 40) had high or low BPtaxon_split according to the taxa involved (Fig. 7). For instance, the supports for splits 37, 39 and 40 from quartets involving the two outgroup reptilian species (alligator and caiman) were lowest and those involving the other species were much higher, whereas the support for split 34 were lowest in the three extinct flightless birds (moa) and the three tinamous. Split 39 is particularly noteworthy. The BPtaxon_split for any of the taxa other than alligator or caiman is very high. It is clear that these two outgroup species are a major reason for the low BP for this tinamous-moa split in the ML tree (Fig. 6).

3.2.2. Taxon stability

BPtaxon scores were calculated and ranked from the smallest to largest for the 24 taxa (Fig. 8). Alligator and caiman appeared the most unstable taxa, followed by the three tinamous and three moas and then by ostrich. This is consistent with the BPtaxon_split plots (Fig. 7) where three of the four unresolved splits (splits 34, 37, 39 and 40) that show BPtaxon_split varies across the taxa have the smallest BP in alligator and caiman. For the RBIC-taxon (Fig. 8), ostrich was the largest among all taxa and therefore it was least stable under this measure. LSI also picked ostrich as the most unstable taxon, followed closely by alligator and caiman and distantly by the two rhea species. Another measure of LSI is based on the entropy of all quartets involving a taxon (Wilkinson, 2006). Using this entropy-based LSI alligator and caiman were estimated to be most unstable followed closely by ostrich. Similarly, TII also found alligator and caiman most unstable, followed by ostrich. Like RBIC-taxon and LSI, the pruned-tree RF distance metric selected ostrich (RF = 3.6) to be most unstable, followed very distantly by chicken and brush turkey (both RF = 5.12) and the other taxa (RF = 5.18).

It should be noted that, while values of some of the measures like BPtaxon and RBIC-taxon are easily interpretable, none of the five taxon stability measures utilized in Fig. 8 have definable thresholds to determine which taxa are 'rogue'. Rather we ranked these measures to see the relative stability of the taxa. Moreover, dot plots of the rankings can show gaps in the stability measures among the taxa and thus provide an *ad hoc* criterion for determining rogue taxa. The same tactic was also applied in the use of TII for rogue detection (Thomson and Shaffer, 2009). Thus based on the visually apparent gap between dots in the various plots in Fig. 8, BPtaxon would pick alligator and caiman to be rogue taxa and LSI and TII would select alligator, caiman and ostrich to be rogues. The other two methods we proposed (RBIC-taxon and pruned-tree RF distance metric) would choose ostrich to be a rogue taxon. Likewise, the RBIC algorithm in the RogueNaRok program only detected ostrich to be a rogue taxon.

4. Discussion

4.1. Differences between the split-specific measures and previous methods

We have introduced several measures based on the distribution of quartets associated with the splits in a tree of interest (e.g., a ML tree or a generating tree) and a set of bootstrap trees to quantify stability at the internal edges (BPsplit and BPtaxon_split) or leaf nodes (BPtaxon and RBIC-taxon). BPsplit is a split-specific quartet measure while the conventional BP is not. We have mathematically proved that BPsplit is always larger than, or equal to, BP (see Eq. (1)). Both the SimuLBA and ratites data show a number of splits where BPsplit is greater than BP. Moreover, the SimuLBA results show that, for internal branches that are not completely resolved (BP < 100%), BPsplit is substantially higher than BP for over half of the correctly estimated branches but it is less often larger than BP for incorrect branches, suggesting BPsplit is a somewhat more robust measure of internal branch support than BP. Both the simulation and ratites data further show BPsplit and BP are strongly correlated with the IC and ICA scores. In many cases the IC (or ICA) and BP are so closely related that the level of the support for the split of interest derived from the IC (the relative BP) is equal to the BP regardless of whether BP and BPsplit are equal or not. However, in at least one case (SimuLBA dataset 7) the relative BP for the split (A1–A4 & B1–B4) inferred from the IC score appears much closer to BPsplit than to BP (Fig. 4B).

At the taxon level, some previous measures, such as the LSI and its extensions as implemented in p4 (Foster, 2004; Mark Wilkinson, personal communication), also use the distribution of quartets. However, they are not split-specific because they consider all 4-taxon groups (permutations of 4 taxa in a tree) rather than those directly associated with a split. Likewise the TII is not a split-specific measure. Furthermore, we believe that the restriction to sets of four taxa in the four subgroups adjacent to the edge is important. For a given split, $S|S'$, sets of four taxa that have two taxa in S and two in S' but don't satisfy the four subgroup restriction are relevant to sums of edges rather than edges. Including such sets of four taxa can thus artificially inflate apparent support for an edge in a tree and, since the numbers of such groups varies across edges, relative support for the edge of interest.

This methodological difference may explain the slightly different estimations of the relative taxon stabilities between the split-specific measures (BPtaxon and RBIC-taxon) and the LSI or TII for the ratites (Fig. 8). For the SimuLBA data all methods predicted the same least stable taxa (Along and Clong) in (nearly) all cases. The additional pruned-tree RF distance measure we propose here is not split-based. The least stable taxa predicted by the latter

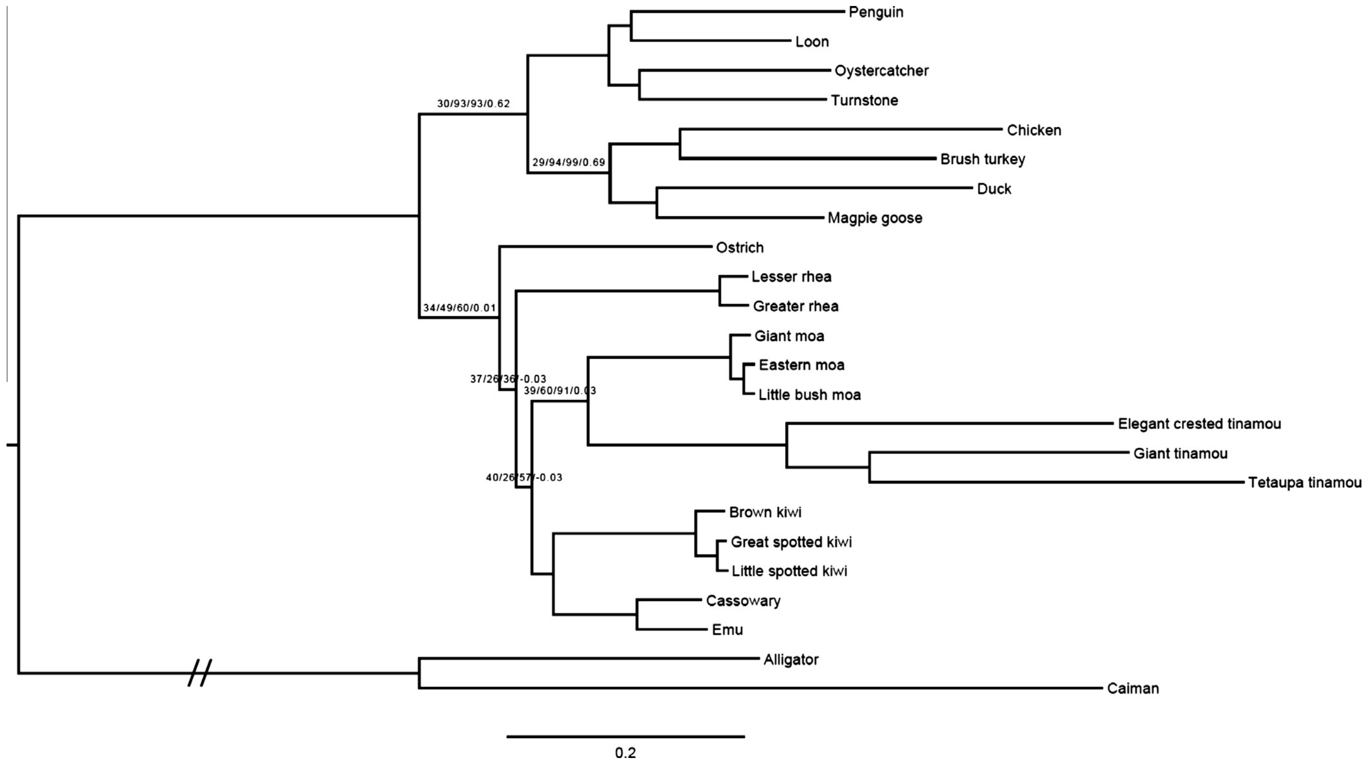

Fig. 6. An estimated ML tree for the 24-taxa ratites data with those branches that have BP less than 95% labeled, on which the first number is the split number, followed by BP, BPsplit and IC. For the unlabeled splits both BP and BPsplit are 100% and IC = 1.

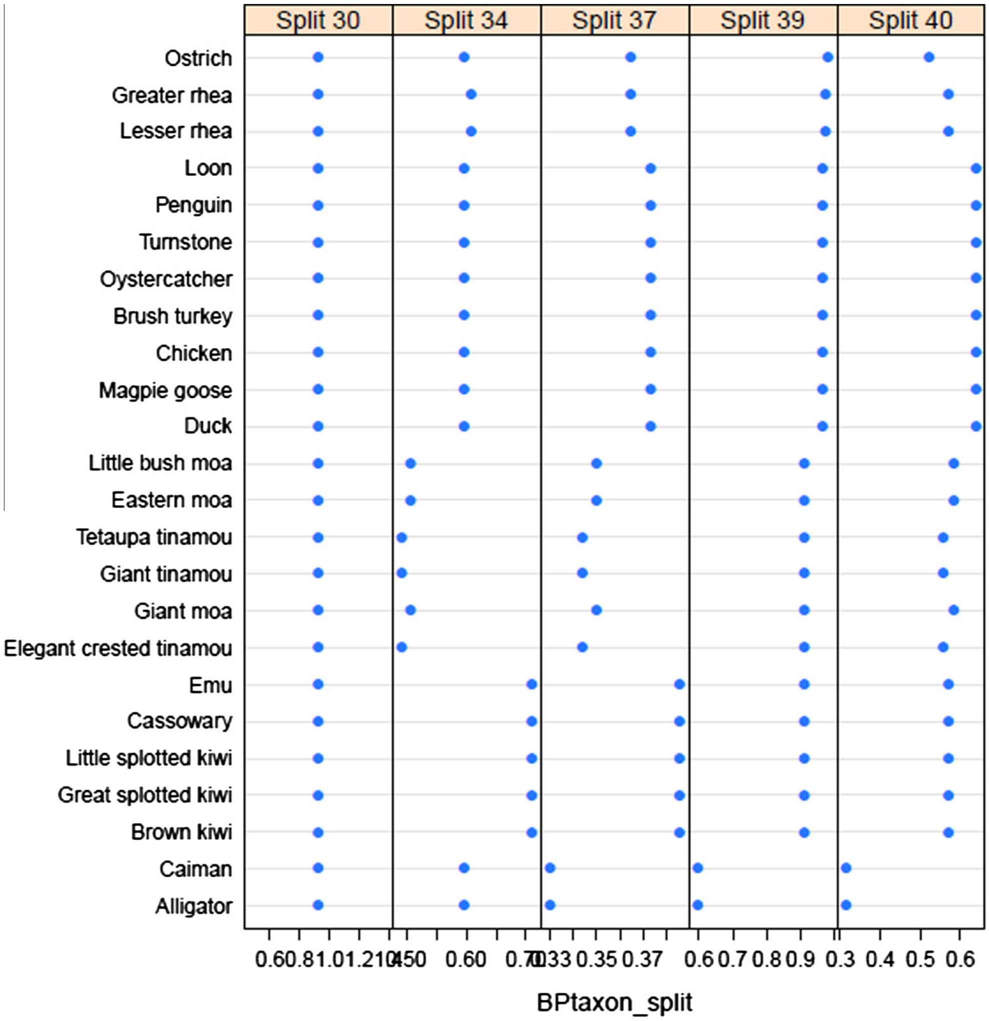

Fig. 7. Dot plots of BPtaxon_split across the 24 taxa for five splits that have BPsplit less than 0.95 in the ML tree of the ratites data. The taxa are ordered according to the values for split 39.

method include Along and Clong for the SimuLBA data and ostrich for the ratites data. They are among the consensus least stable taxa picked by the other measures including BPtaxon, LSI, TII, RBIC-taxon and the RBIC in RogueNaRok. It is consistent with the finding in previous simulations (Aberer et al., 2013) that shows pruning rogue taxa yields consensus trees to be closer to the true tree and the instability of taxa in bootstrap trees reliably predicts taxa for which the position in the ML tree is incorrect.

4.2. The uses of BPtaxon_split

We introduced BPtaxon_split, the average BP for all quartets involving a particular taxon within a particular split. When BP and BPtaxon are identical for a split, BPtaxon_split will be identical across the taxa and its values is the same as BP (see for example split 30 in Fig. 7). When BP and BPtaxon are different for a split, BPtaxon_split will be different among the taxa, a dot plot of which can be used to determine which taxa or taxon group have a large influence on the split (splits 34, 37, 39 and 40 in Fig. 7). For instance alligator and caiman had the lowest support for split 39, which is a crucial split that supports the recent view that ratites are not a monophyletic group, as the flying tinamous form a split with the flightless moas within the ratite groups (Fig. 6 and Phillips et al., 2010).

Although ostrich was detected to be the most unstable taxon with the various taxon stability methods (except BPtaxon), it is not influential for any of the splits shown in Fig. 7. In the cases of split 34 and split 37 the reason is that ostrich is a stand-alone taxon adjacent to these two splits (Fig. 6). As all quartets involving an internal edge having an adjacent terminal edge leading to a stand-alone taxon, must include the taxon on that terminal edge, BPtaxon_split for this taxon cannot deviate substantially from the averages for the other taxa. Therefore, the BPtaxon_split measure was not able to detect the influence of the ostrich on the two splits. In this scenario, an RBIC measure that computes the bootstrap support for a split after removing individual taxa could be useful to detect such an influence, as removing a stand-alone taxon will lead to a fusion of the several edges which will likely increase the BP for the combined edge. However, it remains difficult to know how to compare the impact of such taxa to other taxa, since the removal of no other taxon leads to the fusion of the same two edges in the phylogeny.

In summary, we have proposed several split-specific measures of phylogenetic stability, determined from bootstrap support for quartets for the splits in a phylogenetic tree. BPtaxon quantifies the overall stability of the taxa for a phylogeny and BPsplit measures the stability of internal splits. Although BPsplit is often the same as or similar to BP for many branches in a phylogeny, espe-

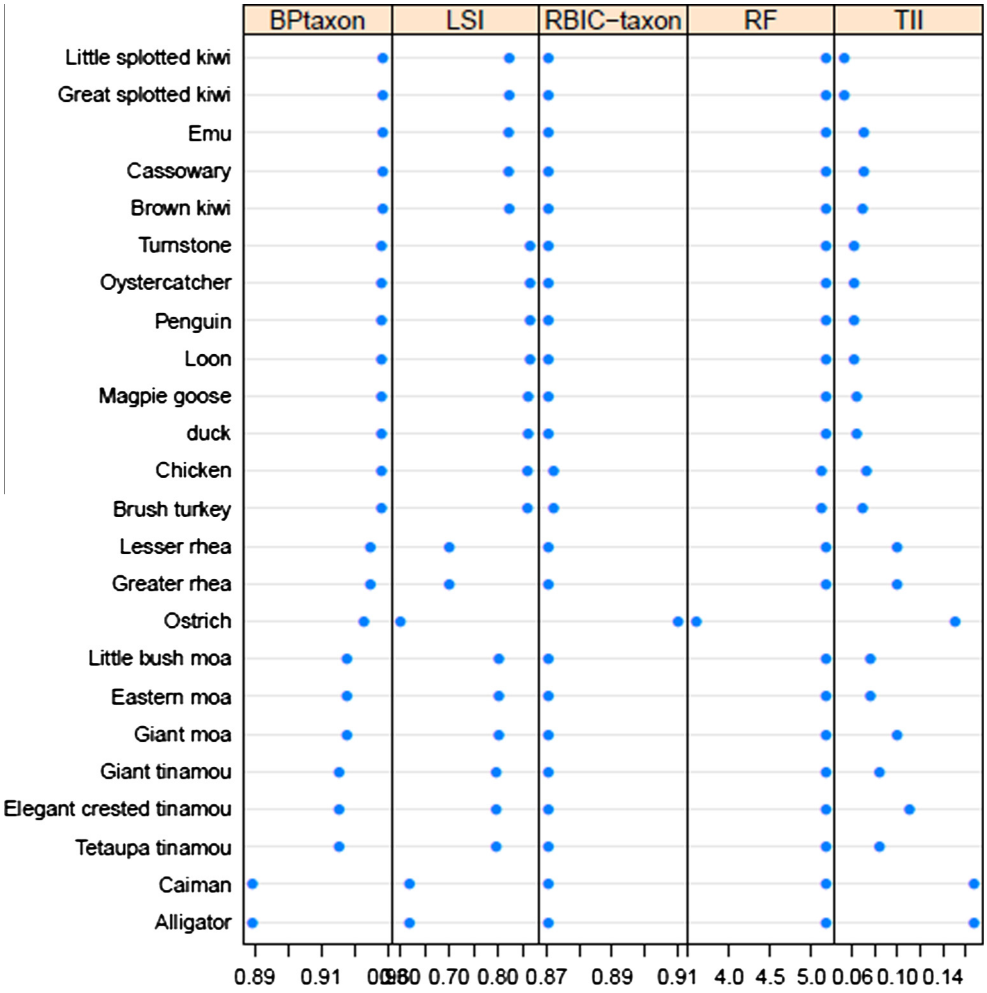

Fig. 8. Dot plots of BPtaxon, LSI, RBIC-taxon, the pruned-tree RF distance and TII across the 24 taxa in the ratites data. The taxa were ranked based on the values of BPtaxon in increasing order. LSI was calculated based on the difference in bootstrap frequencies between the most abundant quartets and the second most abundant quartets involving a taxon. The TII scores have been divided by the number of taxon comparisons.

cially for the strongly supported branches, they can be different. When this is the case, an in-depth analysis of the split with the BPsplit_taxon is helpful, as this measure can show which taxa or groups of taxa have strong, positive or negative, influence on the unresolved splits, thus providing a valuable diagnostic tool to guide taxon sampling in phylogenetic experimental design. Software for the methods is available at http://www.mathstat.dal.ca/~tsusko/.

Acknowledgments

We would like to thank Pablo Goloboff and Mark Wilkinson for very helpful comments in their reviews of the paper. This work was supported by Discovery grants awarded to AJR and ES by the Natural Sciences and Engineering Research Council of Canada (NSERC). AJR acknowledges support from the Canadian Institute for Advanced Research Program in Integrated Microbial Diversity and the Canada Research Chairs program.

References

Aberer, A.J., Krompass, D., Stamatakis, A., 2013. Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. Syst. Biol. 62, 162–166.

Aberer, A.J., Stamatakis, A., 2011. A simple and accurate method for rogue taxon identification. In: Proc. of 2011 IEEE Intl. Conf. Bioinfo. Biomed., pp. 118–122.

Burki, F., Okamoto, N., Pombert, J.F., Keeling, P.J., 2012. The evolutionary history of haptophytes and cryptophytes: phylogenomic evidence for separate origins. Proc. Biol. Sci. B 279, 2246–2254.

Delsuc, F., Brinkmann, H., Philippe, H., 2005. Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. 6, 361–375.

Farris, J.S., Albert, V.A., Källersjö, M., Lipscomb, D., Kluge, A.G., 1996. Parsimony jackknifing outperforms neighbor-joining. Cladistics 12, 99–124.

Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39, 783–791.

Foster, P.G., 2004. Modeling compositional heterogeneity. Syst. Biol. 53, 485–495.

Goloboff, P., Farris, J., 2001. Methods for quick consensus estimation. Cladistics 17, S26–S34.

Goloboff, P., Farris, J., Källersjö, M., Oxelmann, B., Ramírez, M., Szumik, C., 2003. Improvements to resampling measures of group support. Cladistics 19, 324–332.

Goloboff, P., Farris, J., Nixon, K., 2008. TNT, a free program for phylogenetic analysis. Cladistics 24, 774–786.

Goloboff, P.A., Szumik, C.A., 2015. Identifying unstable taxa: efficient implementation of triplet-based measures of stability, and comparison with Phyutility and RogueNaRok. Mol. Phylogenet. Evol. 88, 93–104.

Holland, B., Moulton, V., 2003. Consensus networks: a method for visualising incompatibilities in collections of trees. Alg. Bioinfo. 2812, 165–176.

Huelsenbeck, J., Rannala, B., Masly, J., 2000. Accommodating phylogenetic uncertainty in evolutionary studies. Science 288, 2349–2350.

Huson, D.H., Bryant, D., 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23, 254–267.

Jukes, T.H., Cantor, C.R., 1969. Evolution of Protein Molecules. Academic Press, New York, pp. 21–132.

Maddison, W.P., Maddison, D.R., 2010. Mesquite: A Modular System for Evolutionary Analysis.

Mariadassou, M., Bar-Hen, A., Kishino, H., 2012. Taxon influence index: assessing taxon-induced incongruities in phylogenetic inference. Syst. Biol. 61, 337–345.

Phillips, M.J., Gibb, G.C., Crimp, E.A., Penny, D., 2010. Tinamous and Moa flock together: mitochondrial genome sequence analysis reveals independent losses of flight among ratites. Syst. Biol. 59, 90–107.

Pol, D., Escapa, I.H., 2009. Unstable taxa in cladistic analysis: identification and the assessment of relevant characters. Cladistics 25, 515–527.

Rambaut, A., Grassly, N.C., 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13, 235–238.

Robinson, D.R., Foulds, L.R., 1981. Comparison of phylogenetic trees. Math. Biosci. 53, 131–147.

Salichos, L., Rokas, A., 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497, 327–331.

Salichos, L., Stamatakis, A., Rokas, A., 2014. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. Mol. Biol. Evol. 31, 1261–1271.

Sanderson, M.J., Shaffer, H.B., 2002. Troubleshooting molecular phylogenetic analysis. Ann. Rev. Ecol. Syst. 33, 49–72.

Sheikh, S., Kahveci, T., Ranka, S., Burleigh, J.G., 2013. Stability analysis of phylogenetic trees. Bioinformatics 29, 166–174.

Siddall, M.E., 1995. Another monophyly index: revisiting the jackknife. Cladistics 11, 33–56.

Smith, S.A., Dunn, C., 2008. Phyutility: a phyloinformatics utility for trees, alignments, and molecular data. Bioinformatics 24, 715–716.

Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688–2690.

Swenson, K.M., Chen, E., Pattengale, N.D., Sankoff, D., 2011. The kernel of maximum agreement subtrees. LNBI 6674, 123–135.

Thomas, J.H., 2007. Rapid birth–death evolution specific to xenobiotic cytochrome P450 genes in vertebrates. PLoS Genet. 3 (5), e67.

Thomson, R.C., Shaffer, H.B., 2009. Sparse supermatrices for phylogenetic inference: taxonomy, alignment, rogue taxa, and the phylogeny of living turtles. Syst. Biol. 59, 42–58.

Thorley, J.L., Wilkinson, M., 1999. Testing the phylogenetic stability of early tetrapods. J. Theor. Biol. 200, 343–344.

Wilkinson, M., 1994. Common cladistic information and its consensus representation: reduced Adams and cladistic consensus trees and profiles. Syst. Biol. 43, 343–368.

Wilkinson, M., 1995. Coping with abundant missing entries in phylogenetic inference using parsimony. Syst. Biol. 44, 501–514.

Wilkinson, M., 1996. Majority-rule reduced consensus trees and their use in bootstrapping. Mol. Biol. Evol. 13, 437–444.

Wilkinson, M., 2006. Identifying stable reference taxa for phylogenetic nomenclature. Zool. Scr. 35, 109–112.

Wilkinson, M., Thorley, J.L., Upchurch, P., 2000. A chain is no stronger than its weakest link: double decay analysis of phylogenetic hypotheses. Syst. Biol. 49, 754–776.

Yabuki, A., Kamikawa, R., Ishikawa, S.A., Kolisko, M., Kim, E., Tanabe, A.S., Kume, K., Ishida, K., Inagki, Y., 2014. *Palpitomonas bilix* represents a basal cryptist lineage: insight into the character evolution in Cryptista. Sci. Rep. 4, 4641.