Computing the Internode Certainty and related measures from partial gene trees.

Kassian Kobert[1], Leonidas Salichos[2], Antonis Rokas[3,4], and Alexandros Stamatakis[1,5]

[1]Heidelberg Institute for Theoretical Studies, Heidelberg, Germany
[2]Yale University, Department of Molecular Biophysics and Biochemistry, New Haven, USA
[3]Vanderbilt University, Department of Biological Sciences, 37235 Nashville, USA
[4]Vanderbilt University Medical Center, Department of Biomedical Informatics, 37232 Nashville, USA
[5]Karlsruhe Institute of Technology, Institute for Theoretical Informatics, Postfach 6980, 76128 Karlsruhe, Germany

Abstract

We present, implement, and evaluate an approach to calculate the internode certainty and tree certainty on a given reference tree from a collection of partial gene trees. Previously, the calculation of these values was only possible from a collection of gene trees with exactly the same taxon set as the reference tree. An application to sets of partial gene trees requires mathematical corrections in the internode certainty and tree certainty calculations. We implement our methods in RAxML and test them on empirical data sets. These tests imply that the inclusion of partial trees does matter. However, in order to provide meaningful measurements, any data set should also contain comprehensive trees.

1 Introduction

1.1 Motivation and related work

Recently Salichos and Rokas [7, 8] proposed a set of novel measures for quantifying the confidence for bipartitions in a phylogenetic tree. These measures are the so-called Internode Certainty (IC) and Tree Certainty (TC), which are calculated for a specific reference tree given a collection of other trees with the exact same taxon set.

The calculation of their scores was implemented in the phylogenetic software

RAxML [8, 10].
The underlying idea of Internode Certainty is to assess the degree of conflict of each internal branch, connecting two internal nodes of a phylogenetic reference tree, by calculating Shannon‘s Measure of Entropy [9]. This score is evaluated for each bipartition in the reference tree independently. The basis for the calculations are the frequency of occurrence of this bipartition and the frequencies of occurrences of a set of conflicting bipartitions from the collection of trees.In contrast to classical scoring schemes for the branches, such as simple bipartition support or posterior probabilities, the IC score also reflects to which degree the most favored bipartition is contested.

The reference tree itself can, for example, be constructed from this tree set, or can be a maximum likelihood tree for a phylogenetic alignment. The tree collection may, for example, come from running multiple phylogenetic searches on the same dataset, multiple bootstrap runs [2], or running the analyses separately on different genes or different subsets of the genes (as done for example in [4]). While for the first two cases the assumption of having the same taxon set is reasonable, this does frequently not hold for different genes. Gene sequences may be available for different subsets of taxa, simply due to sequence availability or the absence of some genes in certain species.
In this paper, we show how to compute an appropriately corrected internode certainty (IC) on collections of partial gene trees. When using partial bipartitions for the calculation of the IC and TC scores we need to solve two problems. First, we need to calculate their respective adjusted support (analogous to the frequency of occurrence) (Section 2.1). Unlike in the standard case, with full taxon sets, this information cannot be directly obtained. Then, we also need to identify all conflicting bipartitions (Section 2).

1.2 Bipartitions, Internode Certainty and Tree Certainty

We now briefly define the concepts and notation that we will use throughout the paper. Additionally, we formally define internode certainty and tree certainty.

Bipartition Given a taxon set S, a **bipartition** B of S is defined as a tuple of taxon subsets (X, Y) with X, $Y \subset S$ and $X \cup Y = S$, $X \cap Y = \emptyset$. We write, $B = X|Y = Y|X$.

In phylogenetic trees, a bipartition is obtained by removing a single edge from the tree. Let b be an edge connecting nodes n_1 and n_2 in some unrooted phylogenetic tree T. The bipartition that is obtained by removing b is denoted by $B(b)$, which we define as: $B(b) = X(n_1)|X(n_2)$, where $X(n_1)$ and $X(n_2)$ are all taxa that are still connected to nodes n_1 and n_2 respectively, if branch b is removed.

Trivial bipartition We call a bipartition $B = X|Y$ **trivial** if $|X| = 1$ or $|Y| = 1$.

Trivial bipartitions are uninformative, since having only a single taxon in either X or Y means that this taxon is connected to the rest of the tree. This is trivially given for any tree containing this taxon.
Bipartitions with $|X| \geq 2$ and $|Y| \geq 2$ are called **non-trivial**. In contrast to trivial bipartitions, non-trivial bipartitions contain information about the structure of the underlying topology.
Henceforth, the term bipartition will always refer to a non-trivial bipartition.

Sub-bipartition, super-bipartition We denote $B_1 = X_1|Y_1$ as a **sub-bipartition** of $B_2 = X_2|Y_2$ if $X_1 \subseteq X_2$ and $Y_1 \subseteq Y_2$, or $X_1 \subseteq Y_2$ and $Y_1 \subseteq X_2$.
The bipartition B_2 is then said to be a **super-bipartition** of B_1.

We also need a notion of compatibility and conflict between bipartitions.

Conflicting bipartitions Two bipartitions $B_1 = X_1|Y_1$ and $B_2 = X_2|Y_2$ are **conflicting/incompatible** if there exists no single tree topology that explains/contains both bipartitions. Otherwise, if such a tree exists, they must be compatible. More formally, the bipartitions B_1 and B_2 are incompatible *if and only if* all of the following properties hold (see for example [1]):

$$\begin{aligned} & X_1 \cap X_2 && \neq \emptyset \\ \wedge\ & X_1 \cap Y_2 && \neq \emptyset \\ \wedge\ & Y_1 \cap X_2 && \neq \emptyset \\ \wedge\ & Y_1 \cap Y_2 && \neq \emptyset. \end{aligned}$$

This definition of conflict and compatibility is valid irrespective of whether the taxon sets of B_1 and B_2 are identical or not.

Internode certainty The **Internode certainty** (IC) score (as defined in [7]) is calculated using Shannon‘s measure of entropy [9]. For a branch b we define $IC(b)$ as follows:

$$IC(b) = 1 + X_{B(b)} \cdot log_2(X_{B(b)}) + X_{B^\star} \cdot log_2(X_{B^\star}), \tag{1}$$

where $B(b)$ is the bipartition induced by removing branch b, and $B^\star$ is the bipartition from the tree collection that has the highest frequency of occurrence and is incompatible with $B(b)$. The terms denoted by X are the relative frequencies of the involved bipartitions. That is,

$$X_{B(b)} := \frac{f(B(b))}{f(B(b)) + f(B^\star)},\ X_{B^\star} := \frac{f(B^\star)}{f(B(b)) + f(B^\star)}, \tag{2}$$

where f simply denotes the frequency of occurrence of a bipartition in the tree set.

For the standard case of IC calculations (without partial gene trees), the frequency of occurrence f is simply the number of observed bipartitions in the tree set. In Section 2.1 we will show how to calculate the support (adjusted frequencies) for bipartitions from partial gene trees. We compute this support using the observed frequencies of occurrence. The support for partial bipartitions can then be used analogously to the frequency of occurrence in Equation 2 for calculating the IC scores.
Similarly to the IC score, Salichos and Rokas [8] also introduced the ICA (**internode certainty all**) value for each branch.

Internode certainty all

$$ICA(b) = 1 + \sum_{B^c \in C(b)} X_{B^c} \cdot log_n(X_{B^c}), \tag{3}$$

where $C(b)$, as defined in [8], is $B(b)$ union with a set of bipartitions that conflict with $B(b)$ and with each other, while the sum of support for elements in $C(b)$ is maximized and n is defined as $n = |C(b)|$. Note that $C(b)$ has a slightly different definition in [7].

Again, the terms denoted by X are the relative support of the bipartitions involved in Equation 3. That is,

$$X_{\hat{B}} = \frac{f(\hat{B})}{\sum\limits_{B^c \in C(b)} f(B^c)}$$

for all involved bipartitions $\hat{B} \in C(b)$.
The set $C(b)$ however is not easy to obtain. In fact, as we show in the following observation, maximizing the sum of supports for elements in $C(b)$ renders the search for an optimal choice of $C(b)$ $NP-hard$.

Observation 1 *Finding the optimal set* $C(b)$ *is* $NP - hard$.

This can easily be seen by considering the related, known to be NP-hard, maximum weight independent set problem [3]. Alternatively, the similarity to the problem of constructing the asymmetric median tree, which is also known to be $NP-hard$ [6], can be observed.

For the maximum weight independent set problem, we are confronted with an undirected graph whose nodes are given weights. The task is then to find a set of nodes that maximize the sum of weights, such that no two nodes in this set are connected via an edge. A reduction from this problem, to finding $C(b)$ is straight-forward. Let (W, E) be an undirected graph with weighted nodes W and edges E. Let $B(b) = xy|vz$. First we introduce one bipartition $xz|vy$ for every node in W, with support equal to the node weight. Then, for every pair of bipartitions where the corresponding nodes in W do not share an edge in E, we add four taxa that are unique to those bipartitions, in such a way that they can never be compatible (consider $\ldots ab|cd \ldots$ and $\ldots ac|bd \ldots$). If we find $C(b)$ for the newly introduced bipartitions, the corresponding nodes yield a maximum weight independent set.

For this reason, the definition of the ICA, used and implemented in [8], which we also use here, does not guarantee $C(b)$ to contain the set of conflicting bipartitions that maximize the sum of support. Instead $C(b)$ is constructed via a greedy addition strategy.

Additionally, Salichos and Rokas [7] advocate to use a threshold of 5% support frequency for conflicting bipartitions in $C(b)$. That is, $C(b)$ may only take elements $\hat{B}$ that have support

$$f(\hat{B}) \geq 0.05. \tag{4}$$

This is done to speed up the calculation. Under this restriction, the problem of maximizing the support for $C(b)$ is no longer $NP-hard$. However, the search space is still large enough to warrant a greedy addition strategy, over searching for the best solution exhaustively.

Furthermore, if $B(b)$ does not have the largest frequency among all bipartitions in $C(b)$, the $IC(B)$ and $ICA(b)$ score are multiplied with -1 to indicate this. This distinction is necessary since we may have $|ICA(\hat{b})| = |ICA(b)|$ for some $\hat{b} \in C(b)$. So an artificial negative value denotes that the bipartition in the reference tree is not only strongly contested, but not even the bipartition with the highest support. This can for example occur when the reference tree is the maximum-likelihood tree, and the tree set contains bootstrap replicates.

From the IC scores and ICA scores the respective Tree Certainties TC and TCA can be computed. These are defined as follows:

Tree certainty The TC (**tree certainty**) and TCA (**tree certainty all**) scores are simply the sum over all respective IC or ICA scores. That is,

$$TC = \sum_{\substack{\text{b internal branch} \\ \text{in reference tree}}} IC(b) \tag{5}$$

$$TCA = \sum_{\substack{\text{b internal branch} \\ \text{in reference tree}}} ICA(b). \tag{6}$$

Furthermore, the *relative* TC and TCA scores are defined as the respective values normalized by the number of internal branches b, that is, branches for which $B(b)$ is a non-trivial bipartition.

As we can see, all we need to calculate the IC, TC, ICA and TCA scores is to calculate $C(b)$ (Section 2) and $f(\hat{B})$ (Section 2.1).

2 Adjusting the Internode Certainty

Now we must consider how to obtain the relevant information, that is the sets C and corrected support f, from partial bipartitions.
First, we formally define the input. We are given a so called reference tree T with taxon set $S(T)$ node set $V(T) \supseteq S(T)$ and a set of branches $E(T) \subset V(T) \times V(T)$ connecting the nodes of $V(T)$. Let $\hat{E}(T) \subset E(T)$ be the set of internal branches. That is, for $b \in \hat{E}$ the bipartition $B(b)$ is non-trivial.
Additionally, we are given a collection of trees $\hat{T}$. From this collection we can easily extract the set of all non-trivial bipartitions Bip. The bipartitions in Bip are used to adjust the frequency of other bipartitions. The taxon sets of the bipartitions in Bip are subsets of, or equal to, $S(T)$. We call a bipartition with fewer than $|S(T)|$ taxa a partial bipartition. A bipartition that includes all taxa from $S(T)$ is called comprehensive or full bipartition. From Bip and the bipartitions in the reference tree, we can construct a set of bipartitions P, for which we will adjust the score.

Figure 1 gives an overview of the steps explained in the following sections.

2.1 Correcting the Support

We aim to measure the support the given set of partial trees $\hat{T}$ (or bipartition set Bip) induces for any of the bipartitions in P. We call this the **adjusted frequency** or **adjusted support**. If Bip and P only contain comprehensive bipartitions, the support for any given bipartition is simply equal to its frequency

of occurrence.
In case of partial bipartitions, some thought must be given to the process. Imagine a comprehensive bipartition $B = X|Y$ in P, and a sub-bipartition D of B in Bip. Even though D does not exactly match B, it also does not contradict it. More so, it supports the super bipartition, by agreeing on a common sub-topology.

We distinguish whether the observed sub-bipartition D from Bip is allowed to support any possible bipartition, even those not observed in Bip and P, or just those we observe in P. There seems to be no clear answer as to which of these assumptions is more realistic. The choice is thus merely a matter of definition.

2.1.1 Support of all possible bipartitions: Probabilistic Support

If we assume that an observed sub-bipartition from Bip supports all possible super-bipartitions, not just those in P, with equal probability, the impact on the adjusted support of each such super-bipartition from P ($C(b)$) quickly becomes negligible. Consider the following example:
Let $B = X|Y \in P$, be a super-bipartition of $D = x|y \in Bip$ with $|X \setminus x| + |Y \setminus y| = k$. That is, B contains k taxa that D does not contain. There are 2^k distinct bipartitions with taxon set $X \cup Y$ that also contain the constraints set by D. For $k = 10$ we already obtain $2^{10} = 1024$. That is, the support of D will only increase (adjust) the support of B by less than one permille. More formally, let R_B be the set of sub-partitions in Bip of the comprehensive bipartition B in P and f_D the support for a partial bipartition D in Bip. Then the adjusted support for B, f_B is

$$f_B = \sum_{D \in R_B} \frac{f_D}{2^{(|S(T)| - n_D)}},$$

where n_D is the number of taxa D, and $|S(T)|$ the number of taxa in the reference tree. We use $|S(T)|$ in this formula, since any bipartition in P is implicitly a comprehensive bipartition. That is, even though we do not explicitly

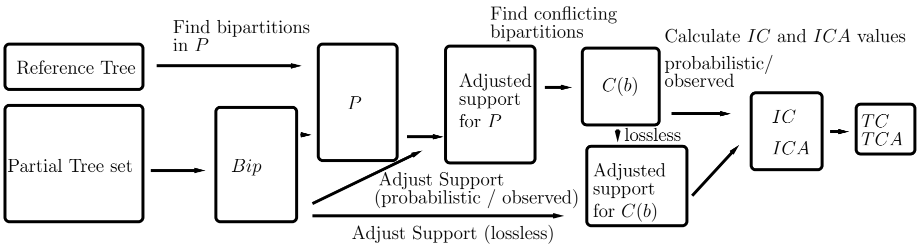

Reference Tree
Partial Tree set
Find bipartitions in P
Bip
P
Adjust Support (probabilistic / observed)
Adjust Support (lossless)
Find conflicting bipartitions
Adjusted support for P
$C(b)$
lossless
Adjusted support for $C(b)$
Calculate IC and ICA values
probabilistic/ observed
IC
ICA
TC
TCA

Figure 1: Overview of the proposed methods.

assign the remaining taxa from a partial bipartition $B = X|Y$ in P to X or Y, they must belong to one of these sets. The missing taxa in D thus have $\frac{1}{2}$ probability to belong to the same set (X or Y) each.

The effect of such an adjustment scheme is that partial bipartitions in Bip with fewer taxa affect the TC and IC scores substantially less than bipartitions with more taxa. This can also be observed in our computational results in Section 3. Since f_B is the sum over the observed frequency times the probability of constructing the actual bipartition implied by B we call this the *probabilistic* adjustment scheme.

2.1.2 Support of observed bipartitions: Observed Support

If we do not want to discard some of the frequency of occurrence when calculating the adjusted support from partial bipartitions, we can distribute their frequency of occurrence uniformly among comprehensive bipartitions in P. When we assume the prior distribution of bipartitions in P to be uniform, this process is simple. For a given partial bipartition D in Bip, with support f_D, let S_D be the set of bipartitions in P that are super-bipartitions of D. Then D contributes $\frac{f_D}{|S_D|}$ support to any $B \in S_D$. In other words, the adjusted support for each full bipartition B is

$$f_B = \sum_{D \text{ s.t. } B \in S_D} \frac{f_D}{|S_D|}. \tag{7}$$

Since this distribution scheme distributes the support for each sub-bipartition among bipartitions that we observed in the tree set only, we call this the *observed* support distribution scheme.

2.1.3 Support of conflicting bipartitions: Lossless Support

One problem with the adjustment strategy explained above is that trees with more taxa typically have more bipartitions in P than trees with fewer taxa. For an intuitive understanding of why this can be problematic consider the example illustrated in Figure 2. Let bipartitions B_1 and B_2 come from the same tree. Further, let bipartition B_3 be the only, and exclusive, sub-bipartition of B_1 and B_2 in Bip. Similarly, let bipartition B_4 be the only super-bipartition of B_5. Let the sub-bipartitions B_3 and B_5 both have a frequency of occurrence of f and let B_1 and B_2 be conflicting with B_4. If we apply the above distribution scheme, bipartition B_1 and B_2 have an adjusted frequency of $f/2$, while B_4 has an adjusted frequency of f. Penalizing bipartitions from trees with larger taxon sets however seems unwarranted. Thus, we propose a correction method that takes this into account. In order to circumvent this behavior we choose

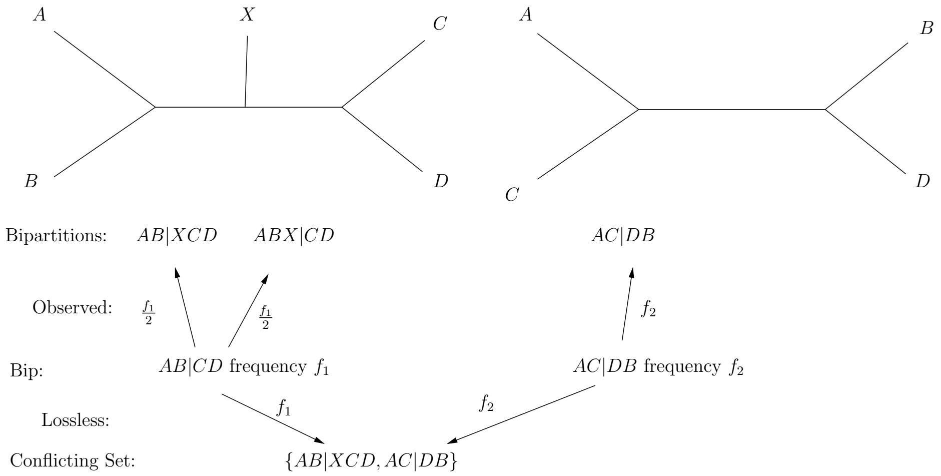


Figure 2: Distribution of adjusted support for observed and lossless adjustment scheme.

to distribute the frequency of any sub-bipartition only to a set of conflicting super-bipartitions (namely bipartitions in $C(b)$). That is:

$$f_B^b = \sum_{D \text{ s.t. } B \in S_D} \frac{f_D}{|S_D \cap C(b)|}. \tag{8}$$

Where S_D is defined as before. Note that, the adjusted support now depends on the set of conflicting bipartitions $C(b)$ which is defined by a branch b. This means that, the adjusted support for a given (conflicting) bipartition must be calculated separately for each reference bipartition $B(b)$.

This distribution scheme distributes the entire frequency of sub-bipartitions only to these conflicting bipartitions. Thus, the sum of adjusted frequencies for all conflicting bipartitions is exactly equal to the sum of frequencies of occurrence of the found sub-bipartitions. For this reason we call this the *lossless* adjustment scheme.

Note that, $C(b)$ is obtained via a greedy addition strategy, depending on the adjusted support of bipartitions. Since the adjusted support according to the lossless adjustment scheme depends on $C(b)$ we obtain a recursive definition. To alleviate this, we simply precompute the above explained probabilistic adjustment scheme to obtain an adjusted support for each bipartition. The set of conflicting bipartitions $C(b)$ is then found with respect to the probabilistically adjusted support values. Then, using $C(b)$, the actual lossless support adjustment is calculated and replaces the probabilistic support in the calculation of IC and ICA values.

For the above example we get the following. Let $\{B_1, B_4\}$ be the set of conflicting bipartitions. Then, the support for B_1 and B_4 after applying the lossless distribution scheme is f for both bipartitions, which is the desired behavior for this distribution scheme.

2.2 Finding Conflicting Bipartitions

From Bip we construct a set of maximal bipartitions P. That is, bipartitions that are not themselves sub-bipartitions of any other bipartition in Bip. Once we have constructed P, we can calculate the internode certainty $IC(b)$ as before. The construction of P is trivial. The set P simply contains all bipartitions that are not themselves strict sub-bipartitions of other bipartitions in Bip. We do this step, since any information contained in a sub-bipartition is also contained in the super-bipartition. That is, the implied gene tree (or species tree) for the super-bipartition can also explain the gene tree for all taxa in the sub-bipartition. How the frequency of occurrence of the sub-bipartition affects the frequency of occurrence of the super-bipartition has been explained in Section 2.1.

We implicitly assume that each bipartition in P should actually contain all taxa from $S(T)$. To achieve this, we keep the placement of the missing taxa ambiguous. That is, we assume that, each missing taxon has a uniform probability to fall into either side of the bipartition.

To construct $C(b)$ greedily as proposed above, the support of the bipartitions must be known. However, the lossless support adjustment scheme explained above is only reasonable on a set of conflicting bipartitions (that is, $C(b)$ itself). To avoid this recursive dependency, we first compute an adjusted support that does not depend on $C(b)$ for this case. (Here we use the so-called *probabilistic adjusted support*, as explained in Section 2.1.1, to obtain an initial adjusted support.) Then, a greedy algorithm is used to approximate the set $C(b)$ with the highest sum of adjusted support, with respect to the initial adjustment. Once $C(b)$ is obtained, the support for all bipartitions in $C(b)$ is adjusted using the new method, which depends on a set of conflicting bipartitions. These new values then replace the initial estimate via the first adjustment scheme.

Keeping the above in mind, we can easily construct $C(b)$ from P for every branch b in $\hat{E}(T)$. Note that, we also defined the reference bipartition $B(b)$ to be in $C(b)$. Thus, we simply start with $B(b)$ and iterate through the elements of P in decreasing order of adjusted support (that is, the probabilistic adjusted support if we are to apply the probabilistic or lossless distribution scheme, and the observed adjusted support if this distribution is desired) and add every bipartition that conflicts with all other bipartitions added to $C(b)$ so far. During this process the threshold given in Equation 4 is applied.

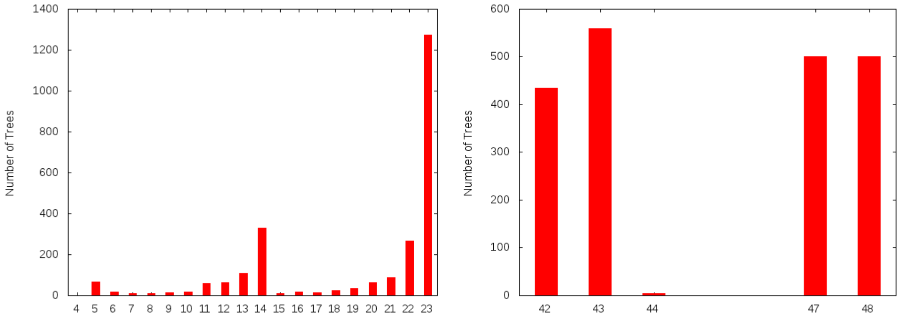
Given $B(b)$, $C(b)$ and Bip we can calculate the IC and ICA values as defined in

Equations 1 and 3 under the *probabilistic* or *observed* adjustment schemes. For the *lossless* adjustment scheme (2.1.3), the actual adjusted frequencies have to be calculated separately for each bipartition in $C(b)$ for all reference bipartitions b in this step.

3 Computational Results

We implemented the methods described in Section 2 in RAxML [10]. In this section we re-analyze the yeast dataset as used by Salichos and Rokas in [7]. The difference is that, we do not only use trees with full taxon sets but also trees with partial taxon sets, which has not been done before. The dataset contains 23 taxa. After applying some filtering techniques we obtain 2494 trees as the basis for our calculations. Of these trees, 1275 contain all 23 taxa. In [7] Salichos and Rokas analyzed a slightly smaller subset of these trees of size 1070. The remaining 1219 trees only contain a partial set of taxa. The number of taxa in these trees ranges from 4 to 22. See Figure 3.a for the distribution of taxon numbers.

Further, we analyze a dataset with 2000 trees and up to 48 taxa. Of these trees, 500 contain the full 48 taxa, the remaining trees contain either 47 taxa (500 trees) or 41-43 taxa (1000 trees). The distribution is illustrated in Figure 3.b. These trees, based on avian genomes, have previously been published in [5].



a) Full tree has 23 taxa. The number of trees with a given amount of taxa are showr

b) Full tree has 48 taxa. The number of trees with a given amount of taxa are shown

Figure 3: Distribution of trees in the trees files.

The software and datasets on which these evaluations are based can be found at `http://www.exelixis-lab.org/material/ICTC.tar.gz`. The probabilistic and lossless distribution scheme are also included in the latest version of RAxML (`https://github.com/stamatak/standard-RAxML`). We chose to omit the implementation for the observed support adjustment from the official RAxML release, as it does not seem to offer any advantages over the other two methods.

We initially report results for the yeast dataset. The results are then confirmed by the second dataset.

First, we assess the effect of including partial gene trees into the analysis. See Table 1 for the results. Table 2 shows the respective results for the ICA values.

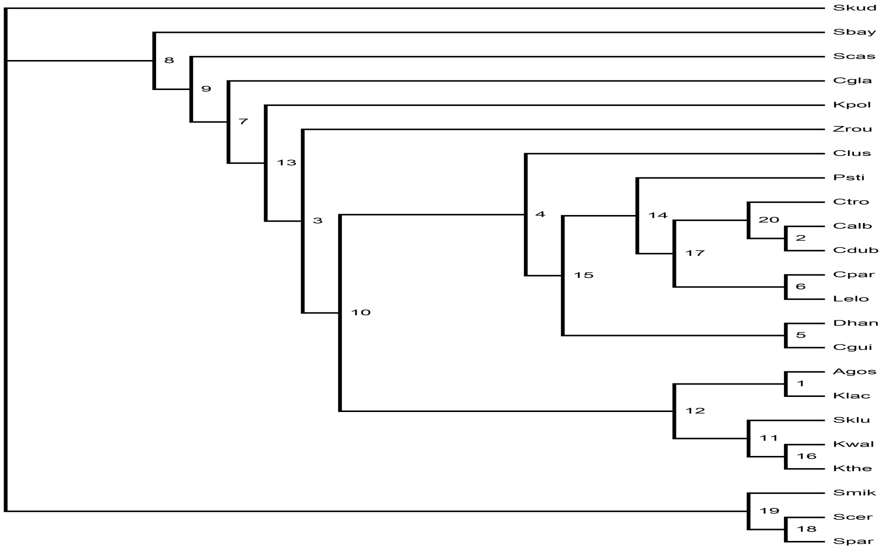


Figure 4: Bipartition numbers corresponding to the presented tables.

Trees	Adjustment	Bip. 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Full	None	9	94	1	99	1	71	2	5	95	56	29	3	14	7	<1	95	71	48	27	75
Full and Partial	Probabilistic	8	92	1	92	2	65	2	6	91	52	28	3	15	7	<1	89	70	46	28	72
Full and Partial	Observed	13	92	2	88	2	67	1	4	62	38	13	3	15	7	1	91	71	55	24	72
Full and Partial	Lossless	15	89	3	68	1	56	<1	5	41	15	2	2	10	7	<1	82	65	39	26	61

Table 1: IC scores for all non-trivial bipartitions multiplied by 100 and rounded down. The bipartition labels are shown in Figure 4.

Trees	Adjustment	Bip. 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Full	None	7	94	2	98	8	71	3	14	95	45	23	8	12	7	9	95	71	48	25	75
Full and Partial	Probabilistic	6	92	1	92	7	60	3	14	91	38	21	13	11	25	11	89	70	46	26	72
Full and Partial	Observed	10	92	2	76	8	67	2	12	62	36	17	13	10	25	7	91	71	53	24	72
Full and Partial	Lossless	10	89	8	68	5	49	3	13	46	29	13	7	9	7	5	82	65	39	27	61

Table 2: ICA scores for all non-trivial bipartitions multiplied by 100 and rounded down. The bipartition labels are shown in Figure 4.

We see that for both, the IC and ICA values, the values for the tree set with full trees is closer to those obtained by the probabilistic adjustment, for full and partial trees combined, than the lossless adjustment scheme for the same tree set. This is expected, since for the probabilistic adjustment, smaller bipartitions contribute less to the overall scores than larger bipartitions. Full bipartitions/trees are thus affecting the outcome most.

The values for the individual IC and ICA scores *can* be higher for the lossless adjustment scheme than for the probabilistic adjustment scheme and the observed adjustment scheme. However, the relative TC and TCA values suggest, that the lossless adjustment attributes a lower certainty to individual bipartitions as well as the entire tree. The actual values are 0.298 for the relative TC score and 0.322 for the relative TCA score for the lossless adjustment; 0.389 and 0.399 for the probabilistic adjustment; and 0.358 and 0.379 for the observed adjustment scheme.

Next, we analyze the behavior of the adjustment schemes if *only* partial trees are provided. See Tables 3 and 4.

Trees	Adjustment	Bip. 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Partial	Probabilistic	61	95	43	93	49	83	39	59	85	64	64	58	46	56	47	93	78	72	66	77
Partial	Observed	74	93	<1	65	38	90	1	82	83	21	71	62	1	57	40	92	90	91	89	90
Partial	Lossless	58	88	12	7	12	42	24	32	68	2	24	12	12	43	38	80	49	66	57	54

Table 3: IC scores for all non-trivial bipartitions multiplied by 100 and rounded down. The bipartition labels are shown in Figure 4.

Trees	Adjustment	Node 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Partial	Probabilistic	54	95	34	93	43	83	40	59	85	58	64	51	46	56	45	93	78	72	66	77
Partial	Observed	74	93	20	65	41	90	21	82	83	21	71	62	1	57	36	92	90	91	89	90
Partial	Lossless	58	88	12	22	12	42	24	29	68	2	27	24	11	43	38	80	49	66	57	54

Table 4: ICA scores for all non-trivial bipartitions multiplied by 100 and rounded down. The bipartition labels are shown in Figure 4.

The relative TC (and TCA) that result from these calculations are 0.668 (0.651) for the probabilistic distribution, 0.619 (0.639) for the observed distribution, and 0.394 (0.407) for the lossless distribution scheme. The relative TC and TCA without correction (obtained from the values shown in Tables 1 and 2), that is for trees with full taxon sets, are 0.406 and 0.409. The higher TC and TCA values obtained for the former two adjustment methods suggest that these approaches are not providing the conflicting bipartitions with a sufficient adjusted support, to compare to the reference bipartition. The reference bipartitions always contain 23 taxa for this data set. Now however, no conflicting bipartition can have that many taxa, as comprehensive trees are not included in the above analysis of only partial trees. In order to avoid overestimating the internode certainty, using the lossless adjustment scheme seems reasonable.

We also have a closer look at the adjusted frequencies of individual bipartitions. To this end, we consider the 15 bipartitions with the largest adjusted support after applying the probabilistic adjustment scheme in Table 5. The adjusted frequencies are applied to the dataset with trees with partial taxon sets only. That is, no tree with all 23 taxa is included. There are 1449 bipartitions that are not strict sub bipartitions of other bipartitions, that is, bipartitions in P. In this table we see that, a high support from the tree set does not seem to induce a higher adjusted support. The highest scoring bipartition, bipartition number 1, has an adjusted support of 10.57, while the support from the tree

Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Adjusted Support	10.57	7.90	7.88	7.57	7.24	6.79	6.69	6.69	6.68	6.44	6.30	6.12	6.09	6.06	6.02
Support from Tree set	8	3	8	9	1	3	10	7	2	7	11	2	10	7	8

Table 5: List of the 15 highest scoring bipartitions after adjusting the frequency by the probabilistic scheme. Support from tree set denotes the number of times this exact bipartition was observed in the dataset. The number only serves as a way to reference the bipartitions in the text.

set is 8. Several bipartitions have a higher support from the tree set, notably bipartitions number 7, 11 and 13 with a support of 10, 11 and 10, respectively. However, for all three, the adjusted support is actually lower than the support observed from the tree set. Similarly, bipartitions with a low support from the tree set can still acquire a high adjusted support for themselves. This can be seen clearly for bipartitions 5, 9, and 12, which only appear once (number 5) or twice (numbers 9 and 12) in the tree set, but accumulate an adjusted support such that they are placed among the 15 highest-scoring bipartitions (out of 1449 bipartitions in P).

A similar behavior can be shown for the observed support adjustment scheme. See Table 6 for the 15 highest scoring bipartitions. The three top scoring bipartitions occur in the tree set only once. Also note that, a low support due to the probabilistic adjustment scheme does not seem to indicate a low support from the observed adjustment scheme. See, for example, bipartitions 8 and 11.

Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Observed support	107	21	21	16	15	14	13	12	11	10	10	9	9	9	9
Probabilistic Support	1.47	1.31	1.11	7.88	10.57	7.57	4.93	<0.01	5.7	6.06	0.78	7.9	7.24	3.13	2.34
Support from Tree set	1	1	1	8	8	9	8	1	11	7	1	3	1	4	4

Table 6: List of the 15 highest scoring bipartitions after adjusting the frequency by the observed support scheme. The number only serves as a way to reference the bipartitions in the text and has no relevance to previous numbering of bipartitions.

Next we consider the adjusted support according to the lossless distribution (see Table 7). The adjusted support depends on the reference bipartition and the (greedily chosen) set of conflicting bipartitions. Thus we only display the adjusted support for a representative set of conflicting bipartitions. We choose the bipartition induced by node number 15, as denoted in Figure 4. The set of conflicting bipartitions is chosen greedily with respect to the probabilistic adjusted support. We note two interesting properties. First, bipartitions 1 and 2 demonstrate that two bipartitions can have the same adjusted support when applying the lossless distribution scheme, even if the adjusted support calculated by the probabilistic adjustment scheme is not identical. Similarly, bipartitions 5 and 6 show the opposite behavior. That is, the probabilistic adjustment scheme yields identical values, while the lossless procedure yields different values. As expected, the adjusted support obtained through the lossless distribution is higher than

Number	1	2	3	4	5	6
Lossless Support	35	35	29	11	6	2
Probabilistic Support	7.89	5.22	3.69	1.47	0.5	0.5
Support from Tree set	8	4	5	2	1	1

Table 7: Adjusted support under lossless adjustment, probabilistic adjustment, and support from data set, for a set of six conflicting bipartitions. The number only serves as a way to reference the bipartitions in the text and has no relevance to any previous numbering of bipartitions.

the adjusted support from the probabilistic scheme. For the given example, the order of bipartitions is preserved, independent of which adjusted values are used. It is worthwhile to note that, this is not the case for all sets of conflicting bipartitions.

Analyzing the second dataset with a total of 2000 trees yields similar results. See Table 8 for the TC and TCA values for this dataset. Specifically, we see that, the probabilistic support for analyzing the full data set, of 2000 trees, again gives TC values more closely in accordance with the values obtained for the analysis restricted to the 500 full trees, than the lossless adjustment scheme.

Trees	adjustment	TC	TCA
48 taxa	None	-3.14	-3.17
41-48 taxa	Probabilistic	-2.44	7.72
41-48 taxa	Lossless	-5.05	-1.35
41-47 taxa	Probabilistic	9.34	15.75
41-47 taxa	Lossless	6.01	6.01

Table 8: IC and ICA scores for different subsets of the data set for the probabilistic and lossless distribution schemes.

Of note is, that the tree set does not support the reference tree well (as evident by the negative TC). At the same time, the TCA under the probabilistic adjustment scheme is actually positive. For this data set, the discrepancy can be explained by the fact that the conflicting bipartitions are also not very distinctly supported. That is, the reference bipartition has (almost) no support and the single most supported conflicting bipartition is supported by some value f. If the support for the reference bipartition is small, the internode-certainty will approach -1. At the same time, let there be a second conflicting bipartition. If the adjusted support of this second bipartition is close to f, the ICA for these bipartitions will be close to 0.0. If the reference bipartition is the bipartition with the highest adjusted support in $C(b)$, this effect is less pronounced.

For the analysis of partial bipartitions only, we again see that the conflicting bipartitions are not as well supported under any tested adjustment scheme. Again, the lossless adjustment scheme yields decreased certainty. Thus the choice of usage of this adjustment scheme is advocated.

From this table, we also see that including trees with the full taxon set seems to yield more reliable certainties than if we restrict the analysis to partial trees only.

4 Conclusion

In conclusion we can say that the inclusion of partial trees into any certainty estimation is beneficial, as the partial trees contain information that is not necessarily contained in the full/comprehensive trees. This is evident by the different TC and TCA scores we obtained.

To calculate the IC and ICA scores, lossless adjustment scheme is found to be most suitable among the tested methods, since it yields more conservative certainty estimates.

Furthermore, we advocate the inclusion of (some) comprehensive trees in any analysis that also includes partial trees to obtain meaningful certainty measurements.

References

[1] D. Bryant. A classification of consensus methods for phylogenies. In M. Janowitz, F.-J. Lapointe, F.R. McMorris, B. Mirkin, and F.S. Roberts, editors, *Bioconsensus*, DIMACS. AMS., pages 163–184, 2003.

[2] Bradley Efron, Elizabeth Halloran, and Susan Holmes. Bootstrap confidence levels for phylogenetictrees. *Proceedings of the National Academy of Sciences*, 93(23), 1996.

[3] Michael R. Garey and David S. Johnson. *Computers and Intractability; A Guide to the Theory of NP-Completeness.* W. H. Freeman & Co., New York, NY, USA, 1990.

[4] Andreas Hejnol, Matthias Obst, Alexandros Stamatakis, Michael Ott, Greg W. Rouse, Gregory D. Edgecombe, Pedro Martinez, Jaume Baguñà, Xavier Bailly, Ulf Jondelius, Matthias Wiens, Werner E. G. Müller, Elaine Seaver, Ward C. Wheeler, Mark Q. Martindale, Gonzalo Giribet, and Casey W. Dunn. Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proceedings of the Royal Society of London B: Biological Sciences*, 2009.

[5] ErichD Jarvis, Siavash Mirarab, AndreJ Aberer, Bo Li, Peter Houde, Cai Li, SimonYW Ho, BrantC Faircloth, Benoit Nabholz, JasonT Howard, Alexander Suh, ClaudiaC Weber, RuteR da Fonseca, Alonzo Alfaro-Nez, Nitish Narula, Liang Liu, Dave Burt, Hans Ellegren, ScottV Edwards, Alexandros Stamatakis, DavidP Mindell, Joel Cracraft, EdwardL Braun, Tandy Warnow, Wang Jun, MThomasPius Gilbert, and Guojie Zhang. Phylogenomic analyses data of the avian phylogenomics project. *Giga-Science*, 4(1), 2015.

[6] Cynthia Phillips and Tandy J. Warnow. The asymmetric median tree a new model for building consensus trees. *Discrete Applied Mathematics*, 71(13):311 – 335, 1996.

[7] Leonidas Salichos and Antonis Rokas. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature*, 2013.

[8] Leonidas Salichos, Alexandros Stamatakis, and Antonis Rokas. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. *Molecular Biology and Evolution*, 2014.

[9] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27, 1948.

[10] Alexandros Stamatakis. Raxml version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 2014.