Anchored distances for quartet-based estimation of phylogenetic trees and applications to coalescent-based analyses (Supplementary Matrial)

Erfan Sayyari[1] and Siavash Mirarab*[1]

[1]University of California, San Diego, Department of Electrical and Computer Engineering

Contents

List of Figures

List of Tables

*Corresponding author: smirarab@ucsd.edu

1 Supplementary Figures and Tables

1.1 Supplementary Tables

Table S1: **Empirical statistics of simulated Avian and Mammalian datasets.** Model condition $2X$ corresponds to the case where ILS is reduced by increasing the branch lengths (2 times longer), and $0.5X$ represents the case where ILS is increased by reducing the branch lengths (2 times shorter). In the same way, the model condition with $0.2X$ corresponds to the case where ILS is reduced by dividing the branch lengths by five. Average Robinso-Foulds (RF) distances between true gene trees and the model species tree are provided in *AD to species tree*. *# gene trees* shows number of gene trees that are available for the corresponding dataset and ILS. *#base pairs* represents number of base pairs, and *# replicates* shows number of replicates for the corresponding dataset and ILS. In column *Ref.*, the reference paper for each dataset is provided. For the Mammalian with ILS level $0.2X$, *# replicates* 5 and 10 are for the model conditions where *# gene trees* is 3200, and 1600 respectively. Also for the Avian dataset with ILS level $1X$, *# replicates* 10 is only for the model condition with *# gene trees* 2000.

	ILS	AD to species tree	# gene trees	# base pairs	# replicates	Ref.
Mammalian	$2X$	18%	200	500,true	20	[3]
	$1X$	32%	200	500,true	20	[3]
	$0.5X$	54%	200	500,true	20	[3]
	$0.2X$	79%	$100, 200, 400, 800, 1600, 3200$	500,true	$5, 10, 20$	[3]
Avian	$2X$	35%	1000	500,true	20	[4]
	$1X$	47%	$200, 500, 1000, 2000$	500,true	$10, 20$	[4]
	$0.5X$	59%	1000	500,true	20	[4]

Table S2: **Empirical statistics of simulated 11-taxon dataset [2]**. Model condition M1 corresponds to the very low ILS, model condition M2 corresponds to low ILS, model condition M3 shows high ILS, and model condition M4 for very high ILS. *AD* represents average bipartition distance between true gene trees and true species trees, expressed as a percentage. The rest of columns are the same as Table S1

dataset	AD	# base pairs	# gene trees	# replicates	Reference
11-taxon M1	15.5%	$10, 25, 50, 100, 200$	$100, 500, 1000$	50	[2]
11-taxon M2	38.3%	$10, 25, 50, 100, 200$	$100, 500, 1000$	50	[5]
11-taxon M3	66.3%	$10, 25, 50, 100, 200$	$100, 500, 1000$	50	[2]
11-taxon M4	85.0%	$10, 25, 50, 100, 200$	$100, 500, 1000$	50	[5]

1.2 Supplementary Figures

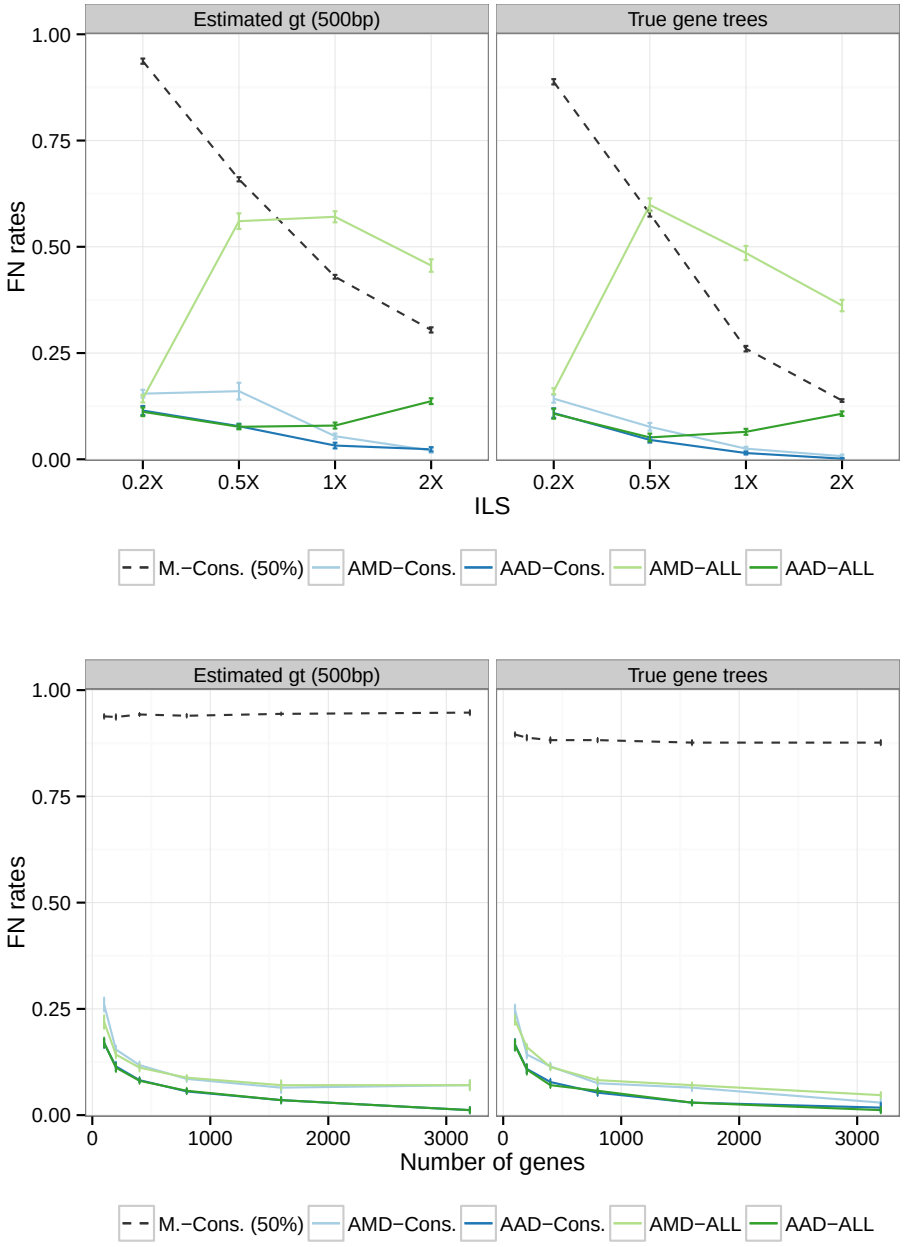

Figure S1: **Accuracy of different implementations of DISTIQUE and Majority Consensus method (50%) for Mammalian dataset.** This figure compares four versions of DISTIUQE on the Mammalian dataset as we vary the amount of ILS, and also shows the missing branch rate for the majority consensus tree. (top) number of genes: 200; (below) ILS: 0.2X. Mean and standard error of species tree error is shown for true and estimated gene trees (500bp alignments). With very high ILS (0.2X), the accuracy for all of the implementations of DISTIQUE are close. With high ILS (0.5X), AAD-Cons, and AAD-ALL have similar accuracy, and AMD-Cons is the next best one. As ILS decreases, when DISTIQUE is applied to the entire dataset, the error goes up, which is more pronounced for AMD-ALL. The results of AMD-ALL for true gene trees is worse than simple Majority Consensus (50%). As discussed before, we attribute this pattern to difficulties of estimating long quartet lengths. When DISTIQUE is used to resolve polytomies in the consensus tree, the accuracy improves with decreased ILS, as expected. Note that even with reduced ILS, the consensus tree on estimated gene trees misses more than 25% of branches, and leaves some polytomies for DISTIQUE to resolve.

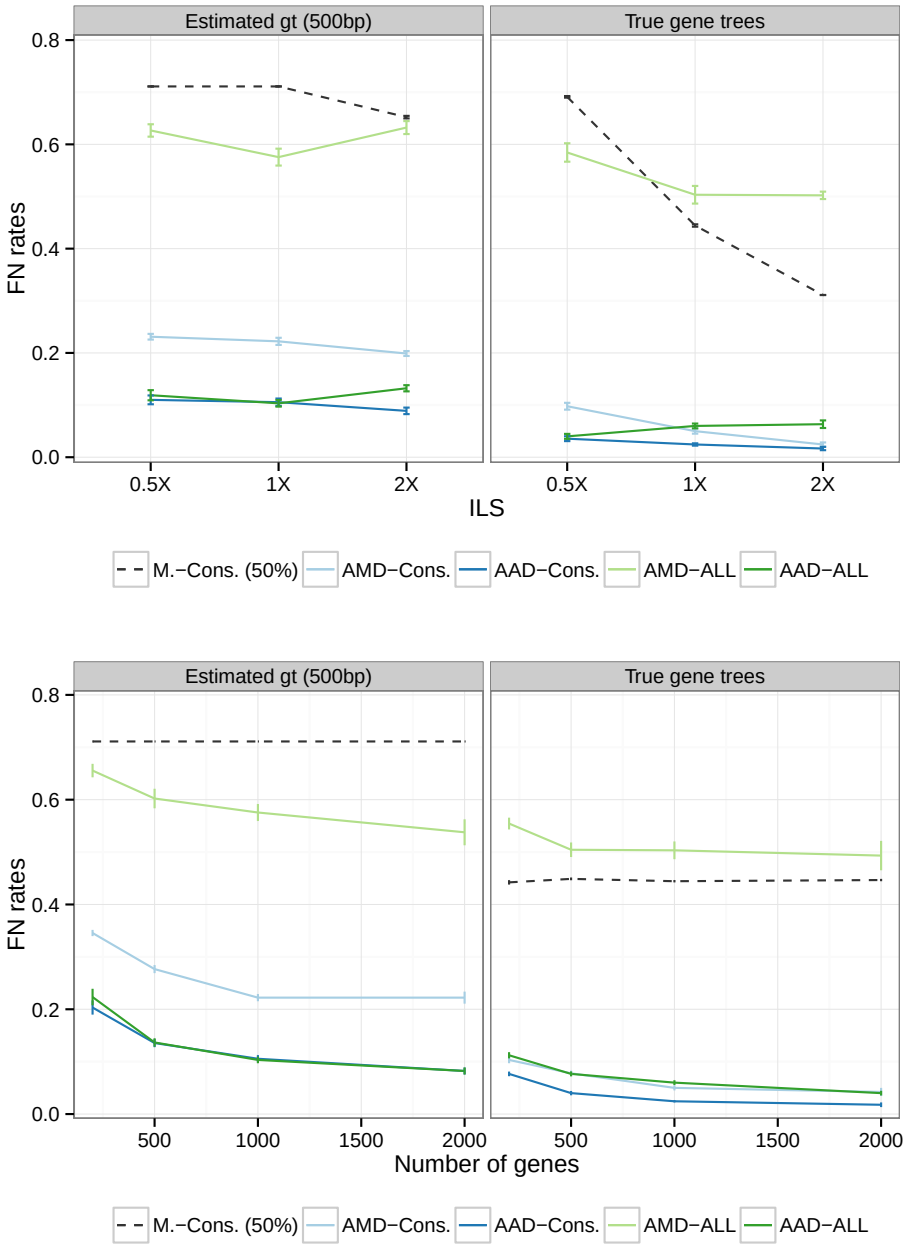

Figure S2: **Accuracy of different implementations of DISTIQUE and Majority Consensus method (50%) for Avian dataset.** This figure compares four versions of DISTIUQE on the Avian dataset as we vary the amount of ILS, and also shows the missing branch rate for the majority consensus tree. (top) number of genes: 1000; (below) ILS: 1X. Mean and standard error of species tree error is shown for true and estimated gene trees (500bp alignments). With high ILS (0.5X), AAD-Cons, and AAD-ALL have similar accuracy, and AMD-Cons is the next best one. As ILS decreases, when DISTIQUE is applied to the entire dataset, the error goes up. The results of AMD-ALL for true gene trees is worse than simple Majority Consensus (50%). As discussed before, we attribute this pattern to difficulties of estimating long quartet lengths. When DISTIQUE is used to resolve polytomies in the consensus tree, the accuracy improves with decreased ILS, as expected. Note that even with reduced ILS, the consensus tree misses more than 30% of branches, and leaves some polytomies for DISTIQUE to resolve.

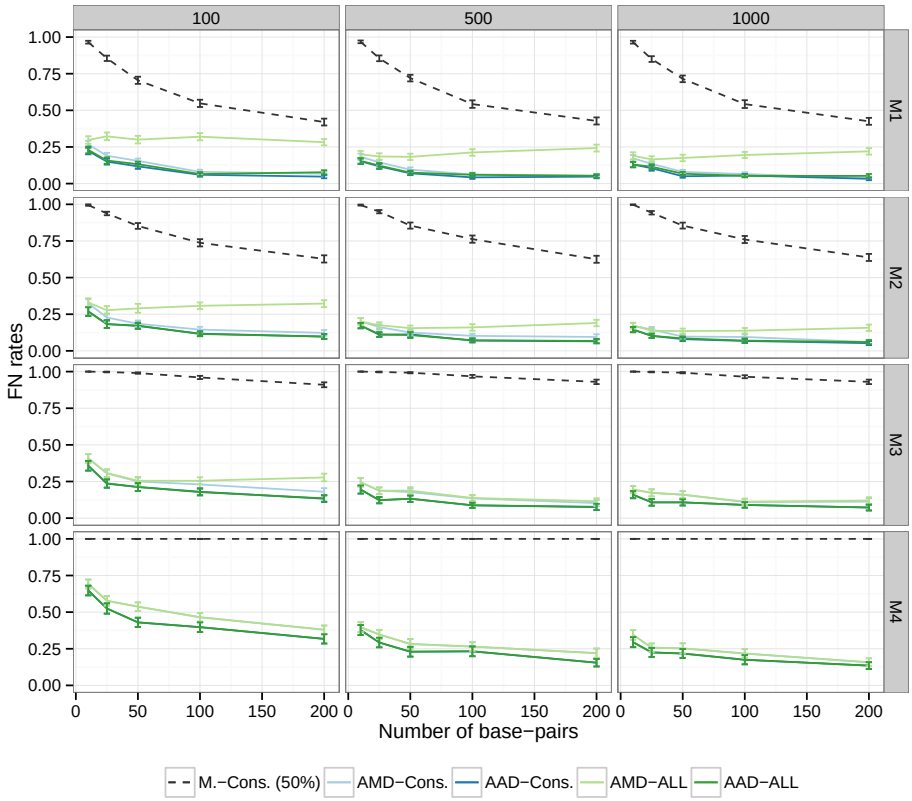

Figure S3: **Accuracy of different implementations of DISTIQUE and Majority Consensus method (50%) for 11-taxon dataset.** This figure compares four versions of DISTIUQE on the 11-taxon dataset as we vary the amount of ILS (rows), and number of genes (columns), and also shows the missing branch rate for the Majority Consensus tree. Mean and standard error of species tree error is shown for estimated gene trees (with number of base pairs varying from 10 to 200). With low and very low ILS (M1 and M2), as number of base pairs increases the missing branch rate of Majority Consensus goes down, but missing branch rate is always above 40%. In high and very high ILS (M3 and M4) Majority Consensus has missing branch rate above 75%, which shows there are big polytomies.

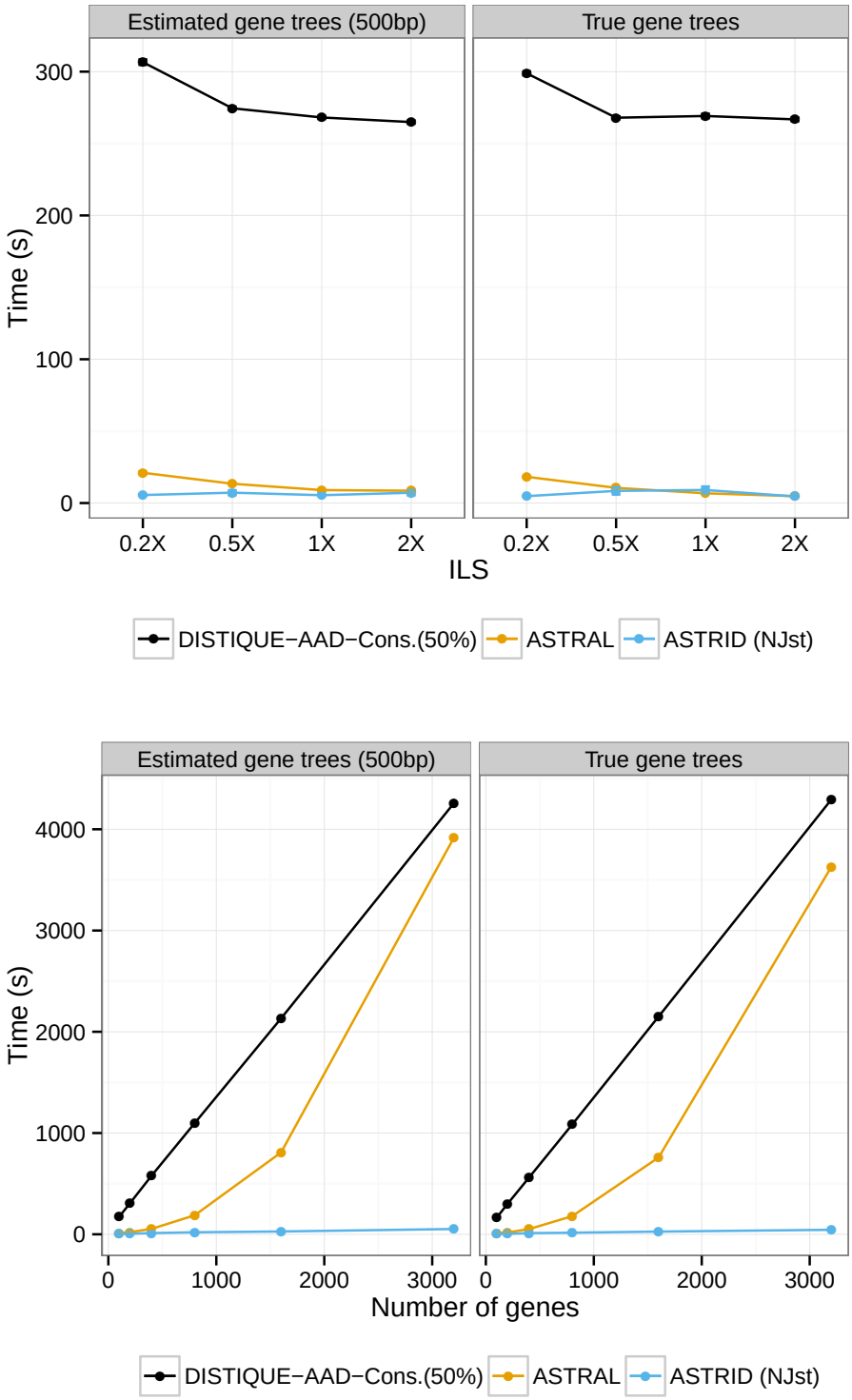

Figure S4: **Running times of DISTIQUE versus other methods for Mammalian dataset.** (top) number of genes: 200; (below) ILS: 0.2X, with 500bp alignments. The time complexity of DISTIQUE and ASTRID (NJst) are almost linear with respect to number of genes, while ASTRAL running time increases with respect to number of genes super-linearly.

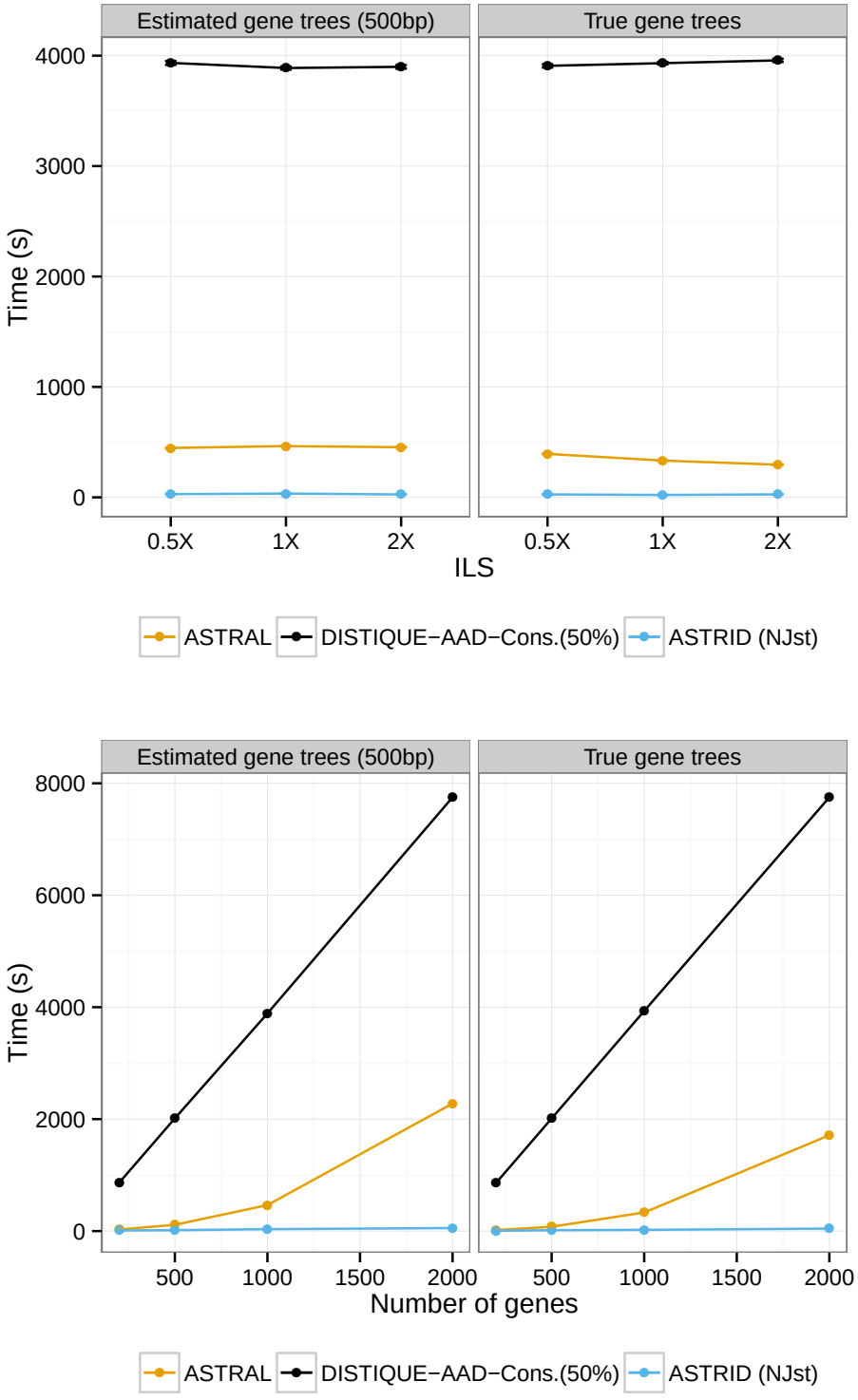

Figure S5: **Running times of DISTIQUE versus other methods for Avian dataset.**. (top) number of genes: 1000; (below) ILS: 1X, with 500bp alignments. This figure shows that the running time of none of the methods depends on ILS level for 1000 gene trees. The running time of DISTIQUE and ASTRID (NJst) are almost linear with respect to number of genes, while ASTRAL running time increases with respect to number of genes super-linearly.

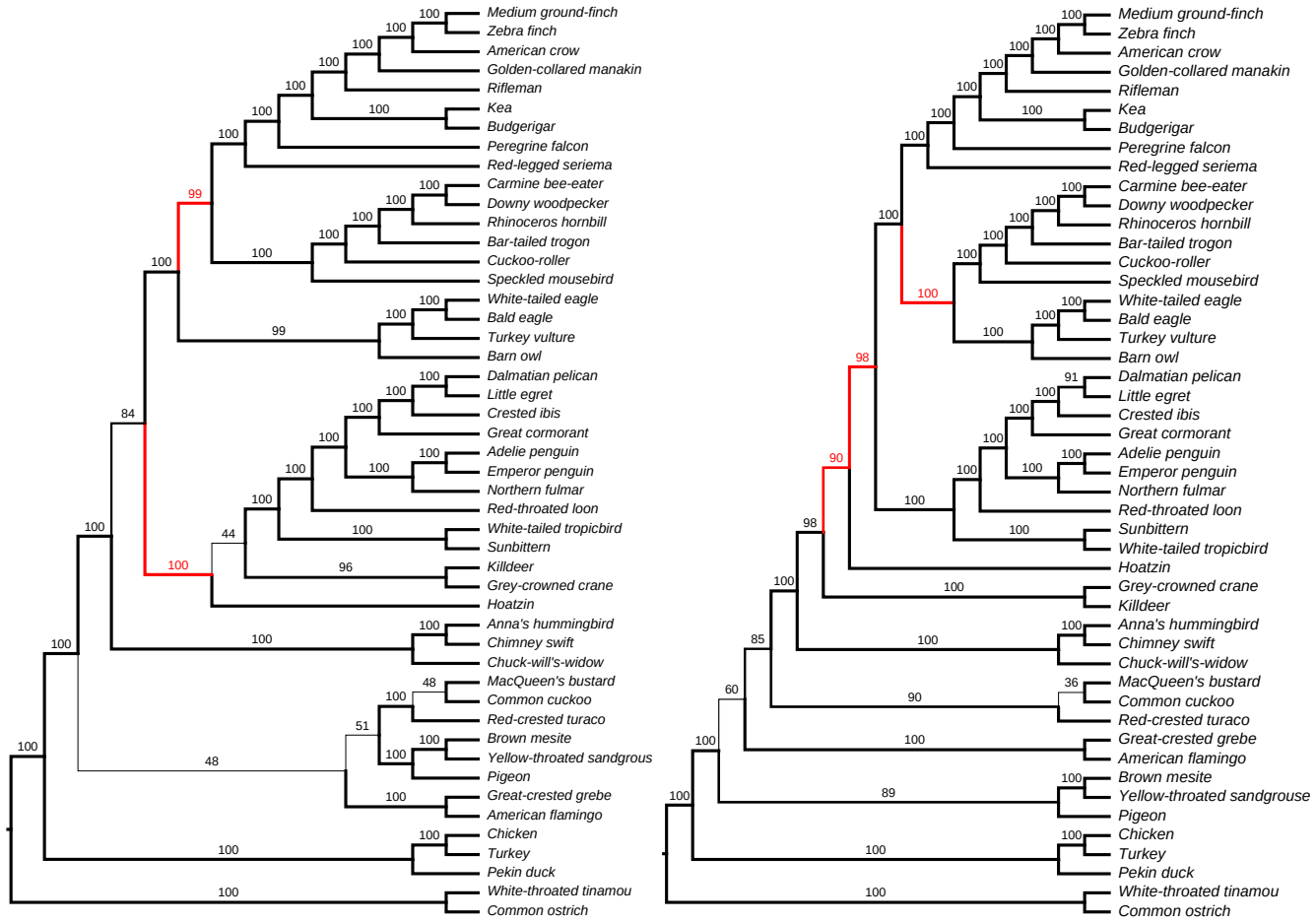

Figure S6: **Species trees generated based on AAD-Cons and ASTRAL using Avian biological dataset [1]**. (left) ASTRAL, (right) AAD-Cons differed in 5 branches, and two of these in DISTIQUE and three of them in ASTRAL had high support, which are represented with red. The presence of highly supported conflict between various coalescent-based methods indicates a need for skeptical and careful analysis of results of any single method on real data.

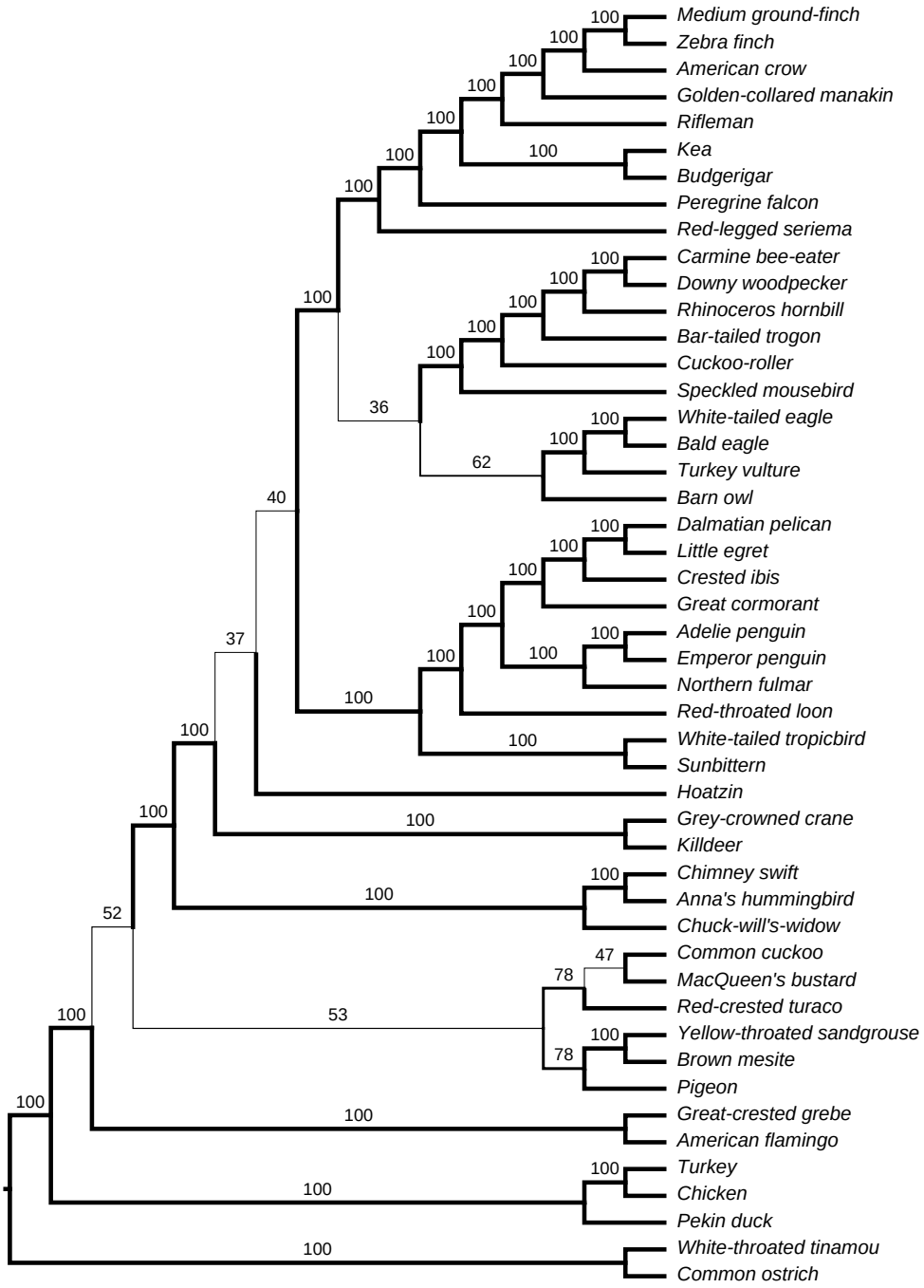

Figure S7: **Species trees generated based on ASTRID (NJst) using Avian biological dataset** DISTIQUE and ASTRID differed on two branches, which both had low support in ASTRID.

2 Difficulties with long branches and need for majority consensus

In this section, we will explain a problematic case that leads us to combine AAD with majority consensus. Figure S8 shows an unrooted species tree, with many long branches, with length L. We assume L is long enough that for our given number of gene trees, with high probability, all gene trees will be topologically identical to the species tree. Note that for any number of genes, there are branches long enough where discordance is highly unlikely (we give one example below). Thus, we assume that for branches of length L, we have zero quartet trees that conflict with them.

In our example, the distance of a to c is L and a to b is $3L$ (note that in our analyses, we only calculate branch length for internal branches). We will show that AAD can be mislead to give a smaller distance for a to b than a to c.

Recall that distances are defined as:

$$D'[a,b] = \sum_{v \in \mathcal{L}-\{a,b\}} D'_v[a,b] = \sum_{u,v \in \mathcal{L}-\{a,b\}} D'_{uv}[a,b] \tag{1}$$

where

$$D'_{uv}[a,b] = \begin{cases} \beta + \alpha . d_D(a,b,u,v) & ab.uv \notin \mathcal{Q}^T \\ \beta - f(d_D(a,b,u,v)) & ab.uv \in \mathcal{Q}^T \end{cases} \tag{2}$$

Also recall that for coalescent-based analyses, $\beta = \ln 3$, $\alpha = 1$, $f(x) = \ln(3 - 2e^{-x})$, and with our use of add-half smoothing, Equation 2 simplifies to

$$D'_{uv}[a,b] = -\ln p(ab.uv) = -\ln\left(\frac{freq(ab.uv) + 0.5}{n + 1.5}\right) \tag{3}$$

where n is the number of genes, and $freq(ab.uv)$ is the number of genes with induced quartet topology $ab.uv$. Using Equations 1 and 3 and considering all selections of anchors u and v, we have

$$\begin{aligned} D'[a,b] =& -\binom{3}{1}.\binom{|X|+1}{1}.\ln\left(\frac{0.5}{n+1.5}\right) - \binom{|X|+1}{2}\ln\left(\frac{n+0.5}{n+1.5}\right) \quad -\binom{3}{2}\ln\left(\frac{0.5}{n+1.5}\right) \\ \approx& -3(|X|+2)\ln\left(\frac{0.5}{n+1.5}\right) \end{aligned} \tag{4}$$

The first term comes from choosing one anchor from $\{u_1, u_2, u_3\}$, and choosing the other anchors from $\{c\} \cup X$. In these cases, a and b are further from each other than the anchors, and because of our assumption about L, the frequency of $ab.uv$ is expected to be zero. The second term comes from choosing both anchors from $\{c\} \cup X$; for these, the frequency of $ab.uv$ is expected to be n. Finally, the last term comes from choosing both anchors from the set $\{u_1, u_2, u_3\}$, where once again, the expected frequency of $ab.uv$ is zero for long enough L, leading to the use of the pseudo count. For large enough n, we can approximate $\ln\left(\frac{n+0.5}{n+1.5}\right) \approx 0$.

The same thing could be written for a and c:

$$\begin{aligned} D'[a,c] =& -\binom{4}{1}.\binom{|X|}{1}\ln\left(\frac{0.5}{n+1.5}\right) - \binom{|X|}{2}\ln\left(\frac{n+0.5}{n+1.5}\right) \quad -\binom{4}{2}\ln\left(\frac{n+0.5}{n+1.5}\right) \\ \approx& -4(|X|)\ln\left(\frac{0.5}{n+1.5}\right) \end{aligned} \tag{5}$$

The first term comes from choosing an anchor from of the $\{u_1, u_2, u_3, b\}$, and the other anchor from X; here, for long enough L, frequency of $ab.uv$ would be expected to be zero. The second term comes from choosing both anchors from X, and the last term comes from choosing both anchors from $\{u_1, u_2, u_3, b\}$; in these cases we expect the frequency of $ab.uv$ to be n.

Comparing the Equations 4 and 5, for $|X| > 2$, the distance between a and b would be smaller than the distance between a and c. This clearly is in contradiction to our tree, so AAD in this case becomes misleading. However, note that all branches with length L are assumed to generate no discordance, and thus will be in the final tree.

Large L: Assume that $L = 16$ (in coalescent units) and we have 1000 gene trees. The probability of having only topologies that agree with the species tree in all 1000 gene trees is $(1-2/3e^{-16})^{1000} = 0.99993$. Thus we expect all n gene trees to have that topology with very high probability. The probability of having only the species tree topology for branches of length $2L$ and $3L$ is even larger.

2.1 Computing pseudo-counts for AMD

In AMD and DISTIQUE-AMD method, in case of zero frequencies for some quartet topologies, without changing the definition of pseudo-count, the relative information about quartets would be lost. For example, assume we have topology $ab.cz$ with long internal branch length, like 16 as mentioned, and $ab.dz$ with internal branch length of 20 (longer than previous length), and no sample for none of the topologies that contradict the species tree. In this case the distance between a, and c is equal to distance of a, and d which is $\ln \frac{0.5}{n+1.5}$, where n is the number of samples. So the relative information is lost. In order to avoid this problem, the definition of pseudo count in AMD and DISTIQUE-AMD is slightly changed. In order to have a pseudo count that could capture the relative distances, first the number of zero quartet topologies are counted. This is called n^0_{ab}. The pseudo count in this case is defined as:

$$\ln \prod_{i=1}^{n^0_{ab}} \frac{0.5}{k_i + 1.5} \tag{6}$$

Where 0.5 comes from our add half estimator, is probability of zero frequencies, and k_i is the number of samples for quartet topology of i.

3 Commands and version numbers

We used ASTRAL version 4.7.8 to find the species trees from gene trees:

```
java -Xmx2000M -jar astral.4.7.8.jar -i [GENE TREES] -o [OUTPUT SPECIES TREE]
```

The ASTRID (NJst) results were produced using the following command:

```
python ASTRID.py -i [GENE TREES] -m [fastme2] -o [OUTPUT SPECIES TREE] -c
[CACHE]
```

CACHE is the distance matrix produced by ASTRID.

For DISTIQUE-AAD we used the following command:

```
python distique-2.py -a [mean] -g [GENE TREES] -m [prod] -o [OUTPUT DIRECTORY]
```

For DISTIQUE-AMD we used the following command:

```
python distique-2.py -a [mean] -g [GENE TREES] -m [min] -o [OUTPUT DIRECTORY]
```

Here the flag a specifies which averaging method to use the partial quartet tables from complete quartet tables around each polytomy.

For comparison and computing false negative missing branchs we used the command available at `https://github.com/smirarab/global/tree/master/src/shell`:

```
compareTrees.missingBranch [-s SPECIES TREE] [-g ESTIMATED SPECIES TREE ]
```

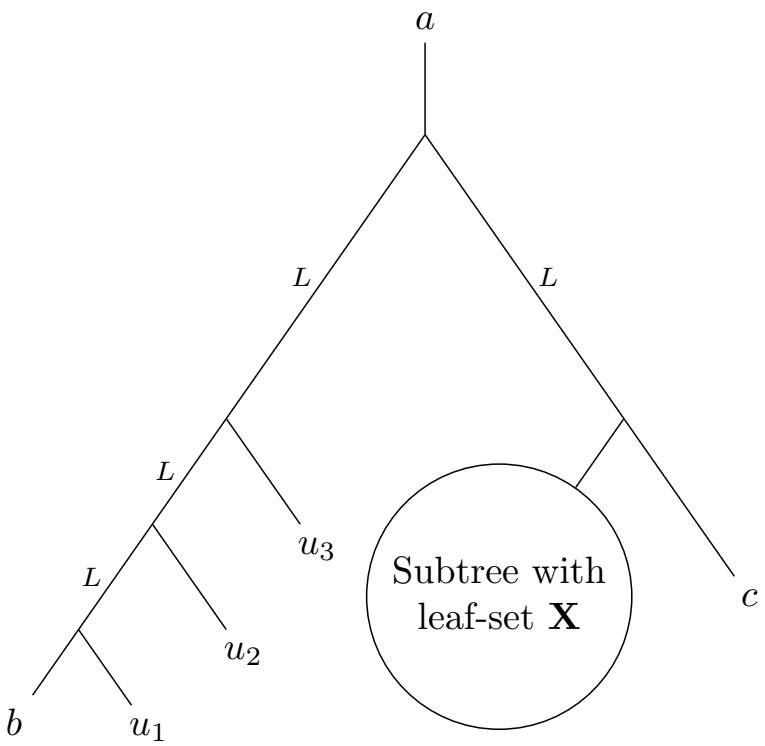

Figure S8: **And example where the AAD method might be misleading for long branches.**

References

[1] Tae Kun Seo. Calculating bootstrap probabilities of phylogeny using multilocus sequence data. *Molecular Biology and Evolution*, 25(5):960–971, 2008.

[2] Jed Chou, Ashu Gupta, Shashank Yaduvanshi, Ruth Davidson, Mike Nute, Siavash Mirarab, and Tandy Warnow. A comparative study of SVDquartets and other coalescent-based species tree estimation methods. *BMC Genomics*, 16(Suppl 10):S2, 2015.

[3] Siavash Mirarab, Md Shamsuzzoha Bayzid, and Tandy Warnow. Evaluating summary methods for multi-locus species tree estimation in the presence of incomplete lineage sorting. *Systematic Biology*, page syu063, 2014.

[4] Siavash Mirarab, Md Shamsuzzoha Bayzid, Bastien Boussau, and Tandy Warnow. Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science*, 346(6215), 2014.

[5] Md Shamsuzzoha Bayzid, Siavash Mirarab, Bastien Boussau, and Tandy Warnow. Weighted Statistical Binning: Enabling Statistically Consistent Genome-Scale Phylogenetic Analyses. *PLoS ONE*, 10(6):e0129183, 2015.