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: **The statistics, and incongruence of true gene trees for 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 (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.

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

Table S2: **Empirical statistics of simulated 11-taxon dataset [3]**. 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	[3]
11-taxon M2	38.3%	10, 25, 50, 100, 200	100, 500, 1000	50	[4]
11-taxon M3	66.3%	10, 25, 50, 100, 200	100, 500, 1000	50	[3]
11-taxon M4	85.0%	10, 25, 50, 100, 200	100, 500, 1000	50	[4]

1.2 Supplementary Figures

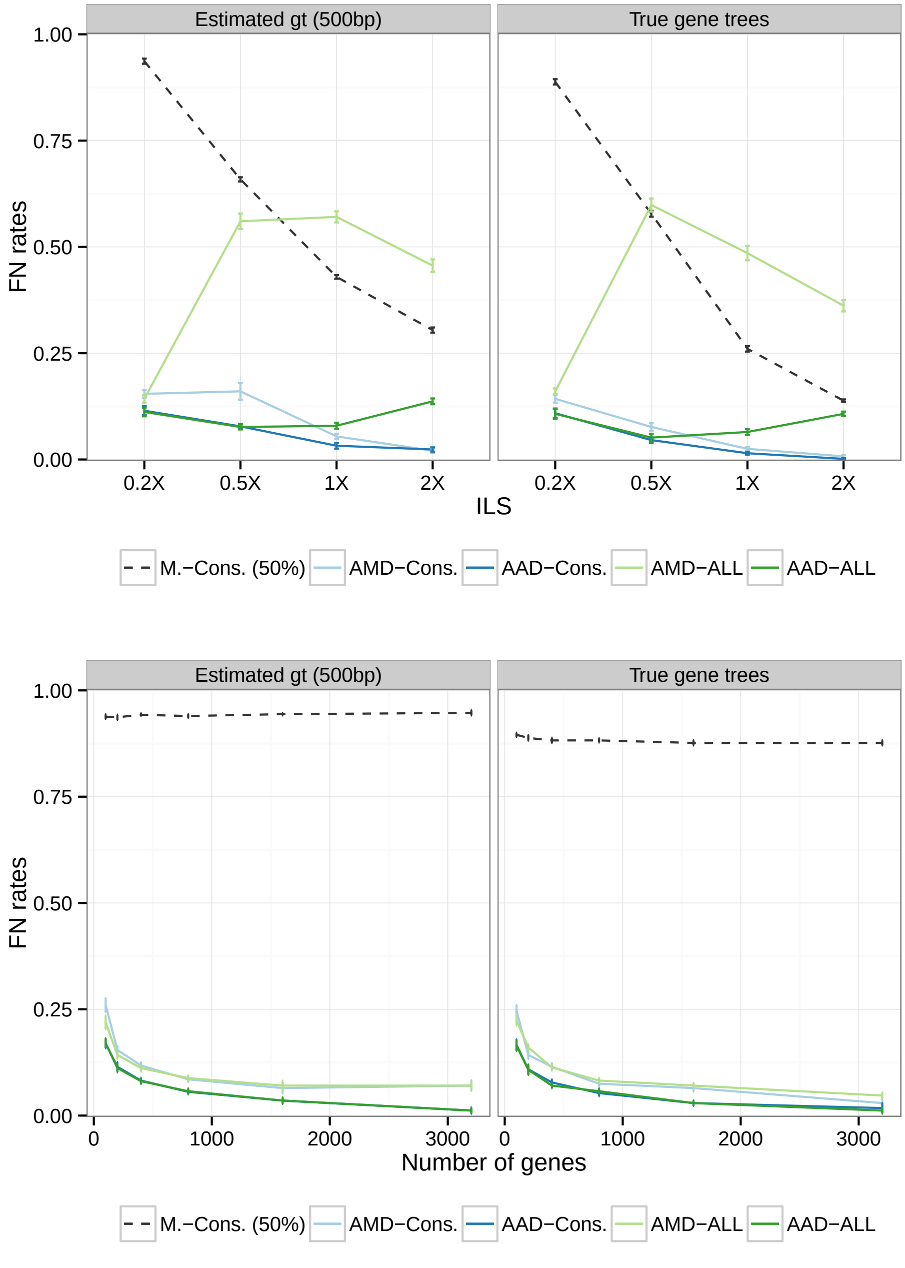

Figure S1: **Different implementations of DISTIQUE versus Majority Consensus method (50%) for Mammalian dataset.** This figure compares four versions of DISTIUQE on the Mammalian dataset as we vary the amount of ILS, and also shows the missing branch rate for the majority consensus tree. (top) number 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), DISTIQUE-AAD, and AAD have similar accuracy, and DISTIQUE-AMD 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. The results of AMD 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.

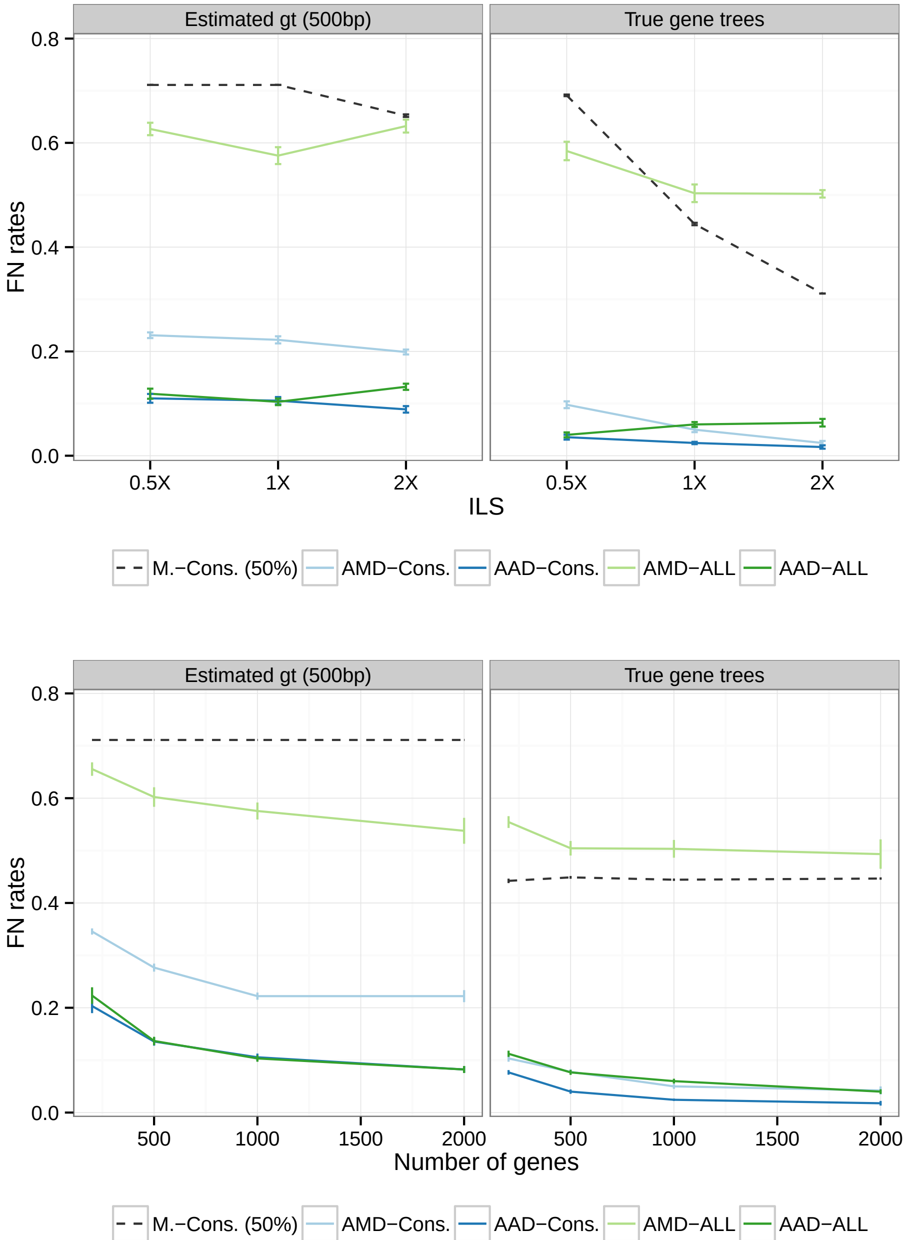

Figure S2: **Different implementations of DISTIQUE versus Majority Consensus method (50%) for Avian dataset.** This figure compares four versions of DISTIUQE on the Avian dataset as we vary the amount of ILS, and also shows the missing branch rate for the majority consensus tree. (top) number of genes: 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), DISTIQUE-AAD, and AAD have similar accuracy, and DISTIQUE-AMD is the next best one. As ILS decreases, when DISTIQUE is applied to the entire dataset, the error goes up. The results of AMD 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.

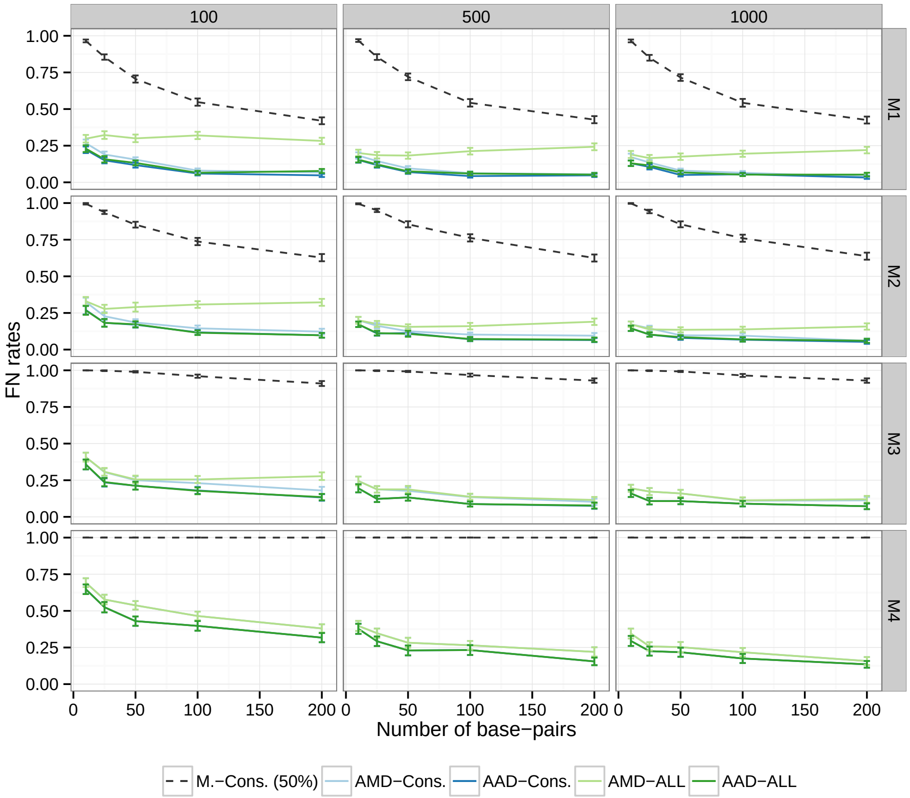

Figure S3: **Different implementations of DISTIQUE versus Majority Consensus method (50%) for 11-taxon dataset.** Columns show different number of genes, and rows show different levels of ILS. Mean and standard error of species tree error is shown for estimated gene trees with varying numbers of sites per gene (50 replicates). In this dataset, no matter what is the ILS level, all versions of DISTIQUE have better accuracy. For low ILS and very low ILS (M1, and M2), as number of base pairs increase the accuracy increases, but it is always above 40%. But for high and very high ILS (M3, and M4), the accuracy of majority Consensus is always above 70%.

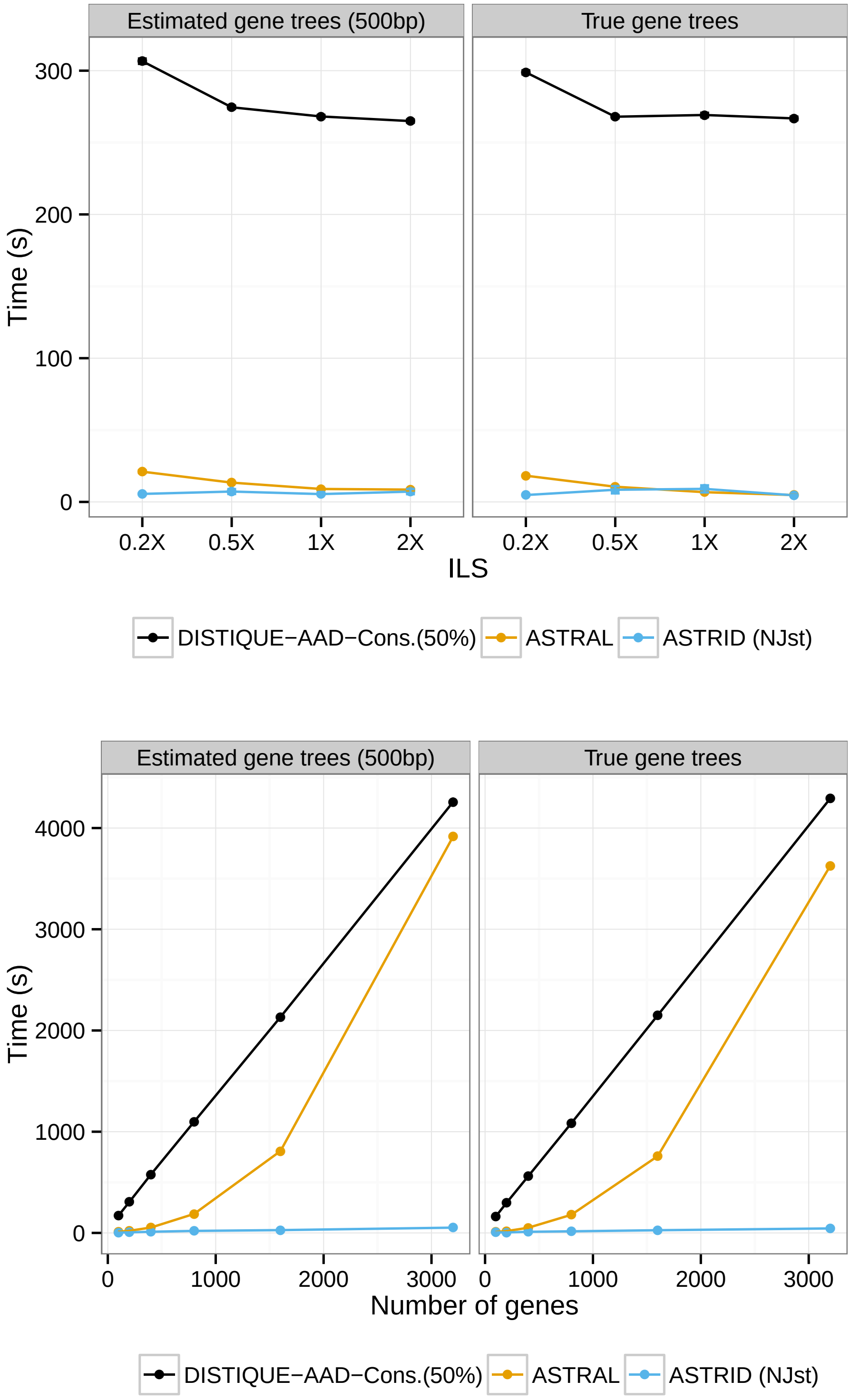

Figure S4: **Running time analyses of DISTIQUE versus other methods for Mammalian dataset.**. (top) number of genes: 200; (below) ILS: 0.2X, with 500bp as alignments. The time complexity of DISTIQUE and ASTRID (NJst) are almost linear with respect to number of genes, while ASTRAL running time increases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

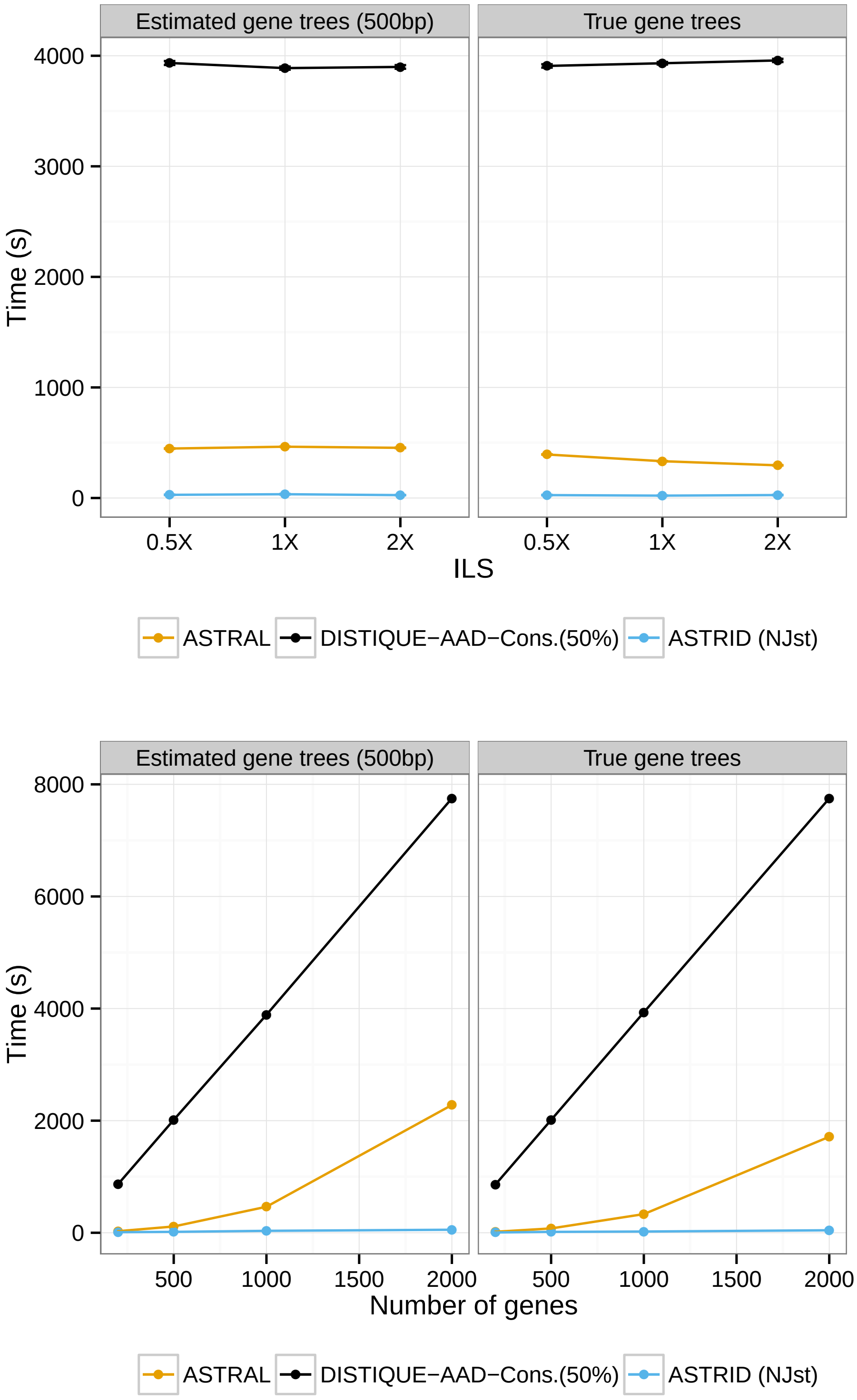

Figure S5: **Running time analyses of DISTIQUE versus other methods for Avian dataset.**. (top) number of genes: 1000; (below) ILS: 1X, with 500bp as alignments. This figure shows that almost none of the methods depend on ILS level for 1000 gene trees. The time complexity of DISTIQUE and ASTRID (NJst)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

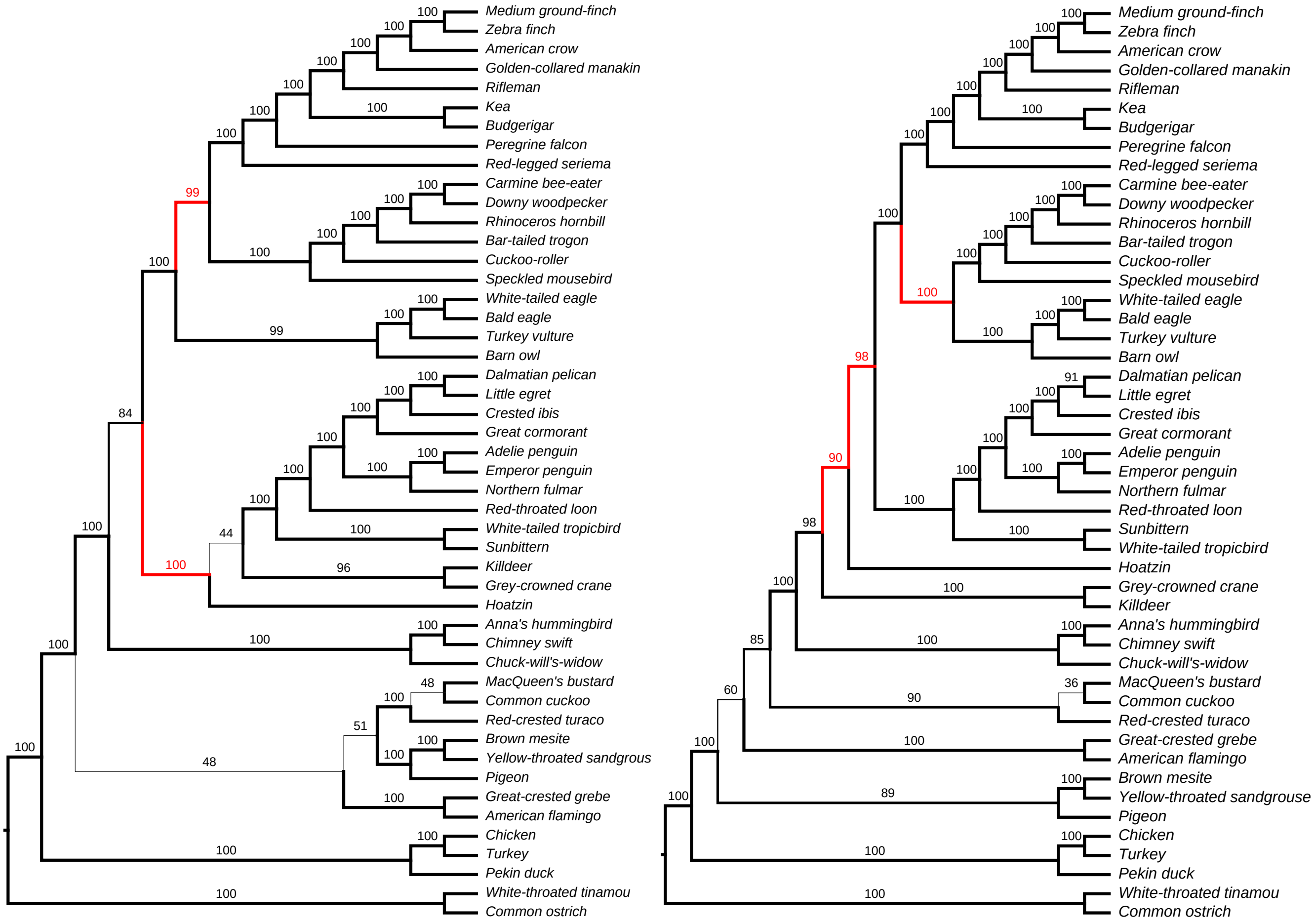

Figure S6: **Species trees generated based on DISTIQUE-AAD and ASTRAL using Avian biological dataset [5]**. DISTIQUE and ASTRAL differed in 5 branches, and two of these in DISTIQUE and three of them in ASTRAL had high support, which are represented with red. The presence of highly supported conflict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

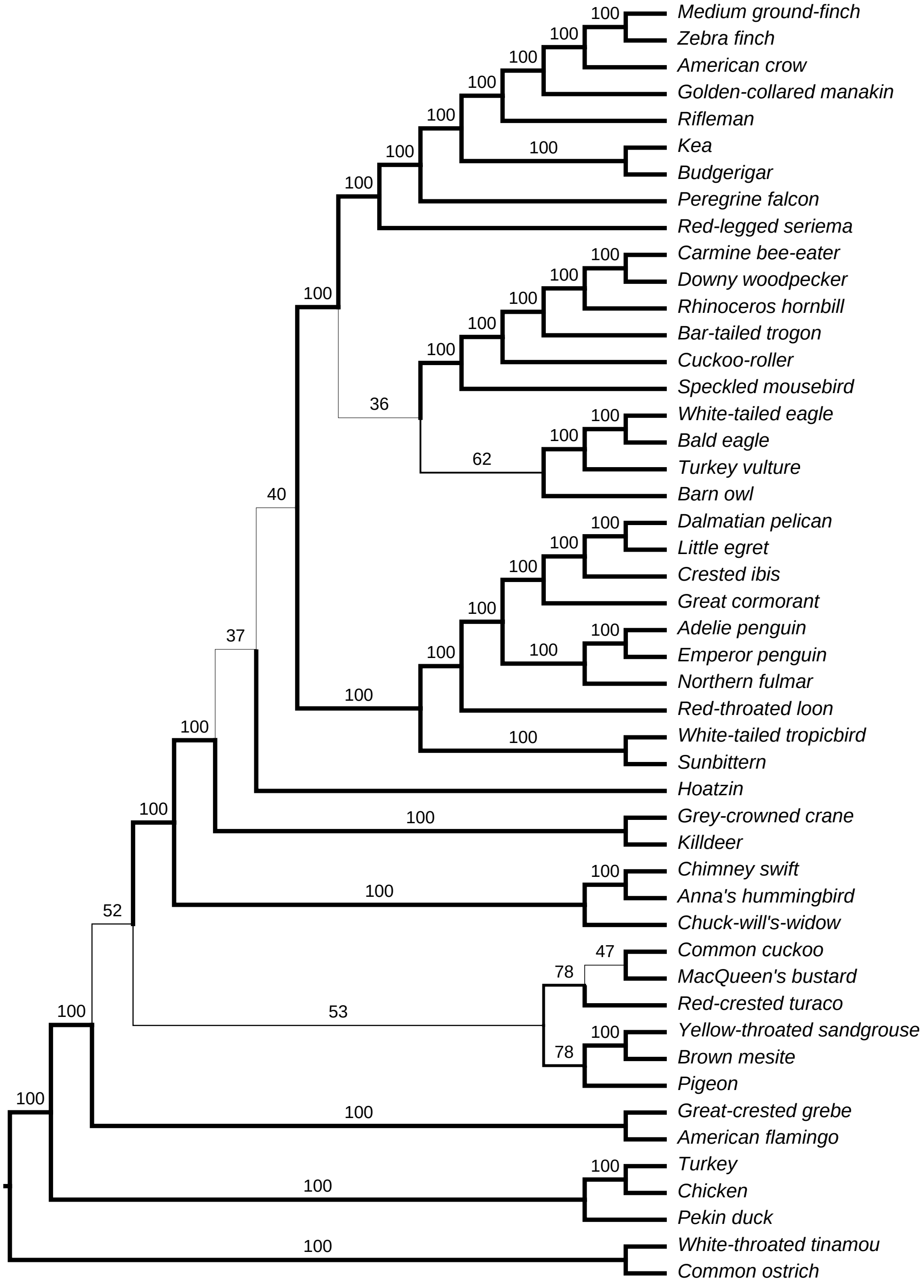

Figure S7: **Species trees generated based on DISTIQUE-AAD and ASTRID (NJst) using Avian biological dataset**. DISTIQUE and ASTRID differed on two branches, which both had low support in ASTRID.

2 Why we need majority Consensus

In this part, we will explain a problematic case that leads us to combine AAD with greedy consensus. Figure S8 is an unrooted tree with a large branch lengths of L. L is not infinity but we will show that even with not so much large L (like 16), in real data getting zero frequencies for topologies that contradicts true topology is highly probable. We assume that we have zero frequencies, so the pseudo-count probability ($\frac{0.5}{n+1.5}$ where n is the total number of samples) for these cases will be used. In true analysis, the taxon a should be closer to taxon c than taxon b. But we will show that without consensus using AAD will be misleading.

The 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

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

Equation 2 is simplified to $D'_{uv}[a,b] = -\ln p(ab.uv)$, where $\beta = \ln 3$, $\alpha = 1$, $f(x) = \ln(3 - 2e^{-x})$, and Equation 2 is simplified to

$$D'_{uv}[a,b] = -\ln p(ab.uv) \tag{3}$$

Using Equation 3 and considering all the cases that a and b are closer to each other topologically than any u and v, the distance between a and b is:

$$\begin{aligned} D'[a,b] = & -\binom{3}{1} . \binom{|\mathbf{X}|+1}{1} . \ln(\frac{0.5}{n+1.5}) + \binom{|\mathbf{X}|+1}{2} 0 \\ & -\binom{3}{2} \ln(\frac{0.5}{n+1.5}) = -3(|\mathbf{X}|+2) \ln(\frac{0.5}{n+1.5}) \end{aligned} \tag{4}$$

The first term comes from choosing one of the u_1, u_2, or u_3, and one from the subset of c union with subtree x. In this case a and b are further from each other so their distance is L from Equation 2. The second term comes from choosing two from the subset of c union with subtree x, and finally the last term comes from choosing two of the u_1, u_2, and u_3. The same thing would be written for a and c:

$$\begin{aligned} D'[a,c] = & -\binom{4}{1} . \binom{|\mathbf{X}|}{1} \ln(\frac{0.5}{n+1.5}) + \binom{|\mathbf{X}|}{2} 0 \\ & + \binom{4}{2} 0 = -4(|\mathbf{X}|) \ln(\frac{0.5}{n+1.5}) \end{aligned} \tag{5}$$

In this case, the first term comes from choosing one of the u_1, u_2, u_3, or b, and one from the subtree x. Since a and c are further from each other, their distance is $-\ln(\frac{0.5}{n+1.5})$ from Equation 2. The second term comes from choosing two from the subtree x, and finally the last term comes from choosing two of the u_1, u_2, u_3, and b.

Comparing the Equations 4 and 5, depending on the $|\mathbf{X}|$, the distance between a and b might be smaller than a and c, so AAD without using consensus in this case might be misleading.

Although ∞ distances in true species trees is impossible, in real analyses there might be topologies of quartets with frequency 0, not because of their distance, but because of lack of samples. This might happen in low ILS cases, with long branches. In order to solve this issue, using majority consensus with reasonable threshold might be helpful.

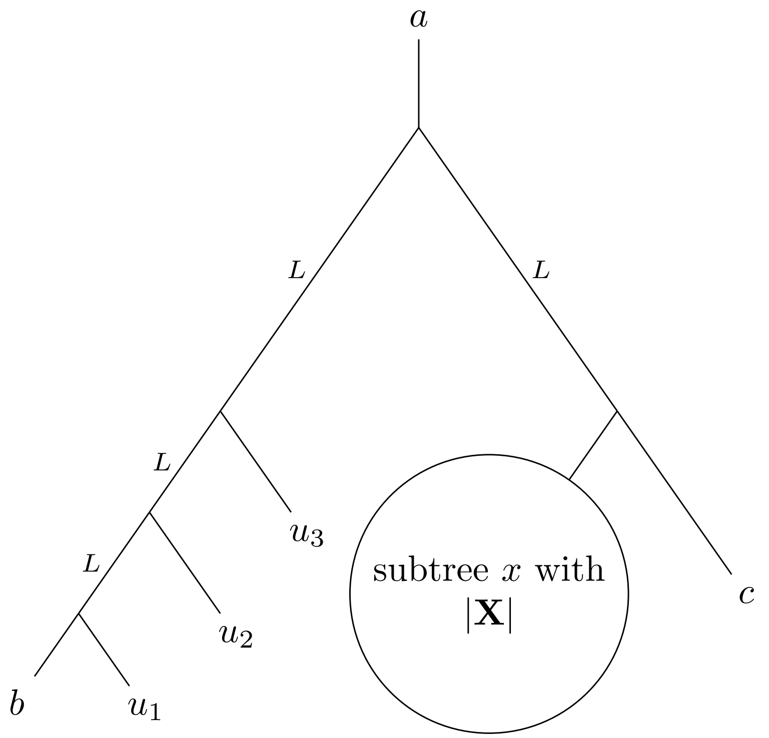

Figure S8: **The counter example that shows without using greedy consensus, AAD method might be misleading.**

To show that this is not a very odd case we assume that the lengths of branches are all 16 and we have 1000 gene trees. The probability of having only topology $ac.bu_3$ in all 1000 gene trees is $(1 - 2/3e^{-16})^{1000} = 0.99992497936$. This probability is also true for the all quartets of $xc.av$ where x is one of the leaves of subtree x, and v is one of the u_1, u_2, u_3, union with b. The probability of having only topology $ac.bu_1$ or $ac.bu + 2$ is even larger than this probability. This shows that even with finite branch lengths, there might be topologies with 0 frequency, and this might be more problematic with real data which translates to infinity distances in our method, and might be misleading.

2.1 Method to compute pseudo counts for AMD

In AMD and DISTIQUE-AMD method, in case of zero probabilities of topologies, without changing the definition of pseudo-count the relative information about quartets will be lost. For example, assume we have topology $ab.cz$ with distance of 10 and $ab.dz$ with distance of 20, 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 ∞. 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 $\ln \prod_{i=1}^{n^0_{ab}} \frac{0.5}{k_i+1.5}$, 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 ]
```

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

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