Supplementary material

Commands
Approximate Bayesian Branch Support Annotation

iqtree2 -s SEQ_ALIGNMENT -te GENE_TREE -m TVM+I+G4 -abayes -pre ANNOTATED_GENE_TREE

Note: When inferring support as a post-processing step, the same model used for inferring the tree should be used, a task that requires care when the original trees are inferred using a different tool (e.g., RAxML). TVM+I+G4 is simply an example.

Running wASTRAL
Exact commands when running on gene trees with approximate Bayesian/Bootstrap/SH-like supports.

astral-hybrid -x 1 -n 0.333 APPROXIMATE_BAYESIAN_ANNOTATED_GENE_TREE
astral-hybrid -x 100 -n 0 BOOTSTRAP_ANNOTATED_GENE_TREE
astral-hybrid -x 1 -n 0 SH_LIKE_ANNOTATED_GENE_TREE
Table S1. Counters \( w^*_w \) are defined for each node \( w \) in each gene tree, and \( Q \) is defined globally. Here, \( X,Y,Z \) are distinct colors of \( A,B,C \). Let \( u,v,r \) be the children of \( w; e \) the parent edge of \( w; p \) be the parent of \( w; P_{2,w} \) be the path between \( x \) and \( w; s(T) = 1 - \prod_{t \in T}(1 - s(t)); m(i,j) = \text{MRCA of } i \text{ and } j \). Counters for leaves are set to zero unless explicitly noted. For each counter, we show a recursive equation on top and the equivalent non-recursive definition on the bottom.

\[
\begin{align*}
    w_X & \equiv \frac{(u_X + v_X) e^{-\ell(T)} \{ \text{for internal node } w; e^{-\ell(T)} \{ \text{for leaf node } w \text{ colored } X} \}}{\sum_{i,e} e^{-\ell(P_{i,e})} \{ \text{for all leaf nodes } i \text{ colored } X \text{ under } w} \}
    \\
    \bigl( w^+_X, w^+_X, w^+_{XY} & \bigr) & \equiv \frac{u_X + v_X + u_X v_X + v_X + u_X v_Y + v_Y}{\sum_{i,e} e^{-\ell(P_{i,e})} \{ \text{for all leaf nodes } i \text{ colored } X \text{ and } Y \text{ under } w} \}
    \\
    \bigl( w^-_X, w^-_X, w^-_{XY} & \bigr) & \equiv \frac{\bigl( (u_X + v_X + u_X v_X + v_X + u_X v_Y + v_Y)(1 - s(e)) \bigr)(u_X v_Y + v_X Y + u_X v_Y + v_Y - u_X) \{ \text{for all leaf nodes } i \text{ colored } X \text{ and } Y \text{ under } w} \}
    \\
    \bigl( w^-_{XY}, w^-_{XY} & \bigr) & \equiv \frac{\bigl( (u_X Y + v_X Y + (u_X - u_X v_X) u_Y + u_Y (v_X - v_X X)) e^{-\ell(e)} \bigr)(u_X Y Z + v_X Y Z + u_Y (v_X Y - v_X X)) e^{-\ell(e)} \{ \text{for leaf nodes } i \text{ colored } X, Y \text{ colored } Y, \text{ and } Z \text{ colored } Z \text{ under } w, \text{ and } m(i,j) \text{ under } m(i,k) \} \}}{\sum_{i,j,k} e^{-\ell(P_{i,j,k})} \{ \text{for leaf nodes } i \text{ colored } X, j \text{ colored } Y, k \text{ colored } Z \text{ under } w, \text{ and } m(i,j) \text{ under } m(i,k) \} \}}
    \\
    \bigl( w^-_{XY} & \bigr) & \equiv \frac{\bigl( u_X Y + v_X Y + (u_X - u_X v_X) u_Y + u_Y (v_X - v_X X)) e^{-\ell(e)} \bigr) e^{-\ell(e)} \{ \text{for internal nodes } w \text{ in } G \} \}}{\sum_{i,j,k} e^{-\ell(P_{i,j,k})} \{ \text{for leaf nodes } i \text{ colored } X, j \text{ colored } Y, k \text{ colored } Z \text{ under } w, \text{ and } m(i,j) \text{ under } m(i,k) \} \}}
\end{align*}
\]

\( Q \equiv \frac{\sum_{G \in G} \sum_{w \in \text{Tree}(G)} W(A | B | C) e^{-\ell(w)}}{\sum_{G \in G} \sum_{w \in \text{Tree}(G)} W(A | B | C) e^{-\ell(w)}} \{ \text{for internal nodes } w \text{ in } G \} \}

\( Q = W[A | B | C] = \sum_{G \in G} W(A | B | C, G) \) (Proposition 5).

Table S2. Running time of species tree inference methods on biological datasets. We use 5.17.3 version of ASTRAL-III if not otherwise clarified.

| Dataset    | n  | k  | Method       | #Cores | Wall-clock time | CPU time |
|------------|----|----|--------------|--------|----------------|----------|
| OneKP      | 1178 | 410 | wASTRAL-h    | 16     | 17.1 min       | 4.57 hr   |
| Canis      | 48  | 44950 | wASTRAL-h   | 1      | 17.2 hr        | 17.7 hr   |
| Avian      | 48  | 14446 | wASTRAL-h    | 16     | 1.76 min       | 28.1 min  |
| Cetacean   | 98  | 3191 | wASTRAL-h    | 16     | 35.2 sec       | 9.39 min  |
| Nominae    | 32  | 853  | wASTRAL-h    | 1      | 8.64 sec       | 8.64 sec  |
| Lepidoptera| 203 | 1930 | wASTRAL-h    | 16     | 2.02 min       | 32.3 min  |
| Papilionidae| 61  | 6405 | wASTRAL-h    | 16     | 1.11 min       | 17.8 min  |
Supplementary Figures and Tables

**FIG. S1.** Species tree error on the S100 dataset with \( k = \{50, 200, 500, 1000\} \) and gene sequence length \{200, 400, 800, 1600\}. Top: comparison of by weighting scheme. Results with aBayes supports are labeled wASTRAL-s and wASTRAL-h; results with bootstrap support are labeled wASTRAL-s* and wASTRAL-h*. Bottom: comparison with other methods, similar to Figure 2a, but with a y-axis kept fixed.
FIG. S2. Species tree error on the S100 dataset (top) and S200 dataset (bottom), similar to Figures 2 and 3 of the main paper, but with y axis kept fixed.
FIG. S3. Lineage Through Time (LTT) plots for three simulated model conditions with $10^{-7}$ (red) and $10^{-6}$ (blue) rates tend to lead to deeper and shallower speciation.
FIG. S4. Species tree error by weighting scheme on the S200 dataset with $k = \{50, 200, 1000\}$ and population size (ILS levels). Species tree shape with parameters E1-6 and E1-7 are used. Results with aBayes supports are labeled wASTRAL-s and wASTRAL-h; results with SH-like support are labeled wASTRAL-s* and wASTRAL-h*.
FIG. S5. Species tree error on the S100 dataset with $k = \{50, 200, 500, 1000\}$ and gene sequence length $\{200, 400, 800, 1600\}$. Results with aBayes supports are labelled wASTRAL-s and wASTRAL-h; results with bootstrap support are labelled wASTRAL-s* and wASTRAL-h*. 
FIG. S6. Species tree error on the S200 dataset with $k = \{50, 200, 1000\}$ and population size (ILS levels). Species tree shape with parameter E1-6 and E1-7 (box columns) and ILS levels (box rows) low ($1e+07$), medium ($2e+06$), and high ($5e+05$) are used. Results with Bayesian supports are labeled wASTRAL-s and wASTRAL-h; results with SH-like support are labeled wASTRAL-s* and wASTRAL-h*.
FIG. S7. ROC of S100 dataset with $k = \{50, 200, 500, 1000\}$ and gene sequence length $\{200, 400, 800, 1600\}$ as we change the threshold of support considered. Results with aBayes supports are labelled wASTRAL-s and wASTRAL-h; results with FastTree-2 bootstrap support are labelled wASTRAL-s* and wASTRAL-h*.
FIG. S8. ECDF of S100 dataset with $k = \{50, 200, 500, 1000\}$ and gene sequence length $\{200, 400, 800, 1600\}$. Results with aBayes supports are labelled wASTRAL-s and wASTRAL-h; results with FastTree-2 bootstrap support are labelled wASTRAL-s* and wASTRAL-h*.
FIG. S9. Binned accuracy-versus-support plot of S100 dataset with $k = \{50, 200, 500, 1000\}$ and gene sequence length $\{200, 400, 800, 1600\}$. Results with aBayes supports are labelled wASTRAL-s and wASTRAL-h; results with FastTree-2 bootstrap support are labelled wASTRAL-s* and wASTRAL-h*.
FIG. S10. ROC of S200 dataset with $k = \{50, 200, 1000\}$ and population size (ILS levels). Species tree shape with parameter E1-6 and E1-7 are used. Results with aBayes supports are labeled wASTRAL-h; results with SH-like support are labelled wASTRAL-h*. 
FIG. S11. ECDF of S200 dataset with $k = \{50, 200, 1000\}$ and population size (ILS levels). Species tree shape with parameter E1-6 and E1-7 (box columns) and ILS levels (box rows) low (1e+07), medium (2e+06), and high (5e+05) are used. Results with aBayes supports are labelled wASTRAL-h; results with SH-like support are labelled wASTRAL-h*. 
FIG. S12. Binned accuracy-verses-support plot of S200 dataset with \( k = \{50, 200, 1000\} \) and population size (ILS levels). Species tree shape with parameter E1-6 and E1-7 are used. Results with aBayes supports are labeled \( \text{wASTRAL-h} \); results with SH-like support are labeled \( \text{wASTRAL-h*} \).
FIG. S13. The distribution of support values of conflicting branches between wASTRAL-h and ASTRAL-III on the 1kp dataset. The ASTRAL-III conflicting branches range between 14% and 99.00% with a mean of 62%. The wASTRAL-h conflicting branches range between 37% and 99.98% with a mean of 78%.
FIG. S14. Inferred species trees (a) from wASTRAL-hybrid with FastTree-2 branch support values as weights using all 459,450 gene trees and (b) from ASTRAL-III using a subset of 100,000 gene trees on canis dataset. Branches support of 100% are omitted.
FIG. S15. Normalized time per round of placement by dividing running time by the total number of rounds of placements for wASTRAL-h on the Canis dataset for various $k$ using the new pipeline.
FIG. S16. Inferred species trees from (a) \textit{w}ASTRAL-hybrid with normalized bootstrap support values as weights and (b) \textit{ASTRAL-III} on gene trees with low (<3\%) bootstrap support branches contracted on avian dataset. Branches support of 100\% are omitted. Branches that disagree with concatenation (blue), MP-EST binned (red) or both (purple) are identified on the \textit{w}ASTRAL-h tree.

FIG. S17. Inferred species trees from (a) \textit{w}ASTRAL-hybrid with normalized Bayesian support values as weights (with clades of taxa from the same species contracted) and (b) \textit{ASTRAL-multi} on cetacean dataset. Branches support of 100\% are omitted. Branches conflicting with RAxML concatenation are marked red.
FIG. S18. (a) RAxML on concatenated genes; (b) wASTRAL-hybrid (top and solid red line) and ASTRAL-III (bottom and dashed red line) on Nomiinae dataset.
FIG. S19. Inferred species trees from (a) wASTRAL-hybrid with normalized bootstrap support values as weights and (b) ASTRAL-III on Lepidoptera dataset.
FIG. S20. Inferred species trees from (a) wASTRAL-hybrid with normalized approximate Bayesian support values as weights and (b) ASTRAL-III on Papilionidae dataset.
FIG. S21. An illustration of the process of creating a random gene tree with branch lengths in SU. Branches in the true species tree $S^*$ are broken into intervals $I_0...I_6$. The species tree with SU branch lengths $S^\dagger$ is created by multiplying each branch length in $S^*$ with a corresponding multiplier; the multipliers are jointly drawn from some distribution and are drawn independently across gene trees. Gene tree $G^*$ is sampled under MSC process from $S^*$ independent of $S^\dagger$. However, it inherits the same division of its lineages into segments as $S^*$ at the same locations. The gene tree with SU branch lengths $G$ is created by translating branch lengths of $G^*$ into SU by multiplying the CU length of each of segment $I_i$ by $\Lambda_{S^\dagger}^{I_i}$, the multiplier associated with the segment $I_i$ in $S^\dagger$ and hence $G$. 
FIG. S22. The species tree estimation error (FN) of wASTRAL-h on S100 dataset as we change the number of rounds of placements in the base algorithm (r). The most difficult case where gene length = 200 and k = 50 is selected. Mean and standard error (50 replicates) are shown in blue.
Supplementary Algorithm
Algorithm S1 Recursive placement algorithm. Place inserts the species $i$ into an existing species tree $S$ and computes tripartition scores $W(A|B)(C,G):=\sum_{C \in G} W(A \cap L_G|B \cap L_G; C \cap L_G, G)$ for all tripartitions resulting from adding $i$ to each branch of $S$. A global counter $Q$ and a set of per-node counters $w_A, w_B, w_C, w_+, w_-, w_{+\cdot}, w_{-\cdot}, w_{\cdot+}, w_{\cdot-}$ are all initialized to 0. OptimalTreeDP is defined in Algorithm S2. Each gene tree is rooted on an arbitrary branch $e$ and the support of $e$ is kept for the branch on one side of the root and zero support is given to the branch on the other side of root. $L_w$ is the set of leaves under $w$.

1: \textbf{procedure} \textsc{Place}(i,S,G) \Comment*[r]{Places species $i$ on tree $S$ according to $G$}
2: \hspace*{1em} $W \leftarrow$ empty lookup table \Comment*[r]{global variables}
3: \textsc{ColorLeafSet}(L_S, C, S, G, W) \Comment*[r]{Color all leaves of $S$ as $C$}
4: \textsc{ColorLeafSet}(\{i\}, B, S, G, W) \Comment*[r]{Color new species $i$ as $B$}
5: \textsc{ColorNode}(the root of $S$, S, G, W) \Comment*[r]{Traverse $S$ bottom up}
6: $O \leftarrow \textsc{OptimalTreeDP}(L_S \cup \{\{\}, L_S \cup \{\{\}, W\})$
7: \textbf{return} $(W, O, \text{edge of } S \text{ onto which } i \text{ is added to get } O)$

8: \textbf{procedure} \textsc{ColorLeafSet}(L', X, T, G, W) \Comment*[r]{Condition: Coloring $L'$ as $X$ should match $T$}
9: \hspace*{1em} for $G \in \tilde{G}$ do
10: \hspace*{2em} $W[\{j\}] \leftarrow \textsc{UpdateCounters}(\text{leaf node corresponding to } j \text{ in } g, X)$
11: \textbf{procedure} \textsc{ColorNode}(w, i, S, G, W) \Comment*[r]{On start: $i$ is $B$, others are $C$; On exit: $w$ is $A$, others kept}
12: \hspace*{1em} if $w$ is a leaf then
13: \hspace*{2em} \textsc{ColorLeafSet}(L_w, A, L_w, [\{\}], L_S - L_w, G, W)
14: \hspace*{1em} else
15: \hspace*{2em} $(u, v) := (\text{the larger child of } w, \text{the smaller child of } w)$
16: \hspace*{2em} \textsc{ColorNode}(v, i, S, G, W) \Comment*[r]{Recurse on $v$, the smaller child}
17: \hspace*{2em} \textsc{ColorLeafSet}(L_v, C, \emptyset, G, W) \Comment*[r]{Undo coloring of $v$ to enable recursing on $u$}
18: \hspace*{2em} \textsc{ColorNode}(u, i, S, G, W) \Comment*[r]{Recurse on $u$, the large child}
19: \hspace*{2em} \textsc{ColorLeafSet}(L_u, B, L_u, [\{\}], L_S - L_u, G, W) \Comment*[r]{Tripartition of $w$ when adding $i$ above $u$}
20: \hspace*{2em} \textsc{ColorLeafSet}(\{\}, A, \{\}, L_S - L_u, G, W) \Comment*[r]{Tripartition of $w$ when adding $i$ above $u$}
21: \hspace*{2em} \textsc{ColorLeafSet}(\{\}, C, L_u, [\{\}], L_S - L_u, G, W) \Comment*[r]{Tripartition of $w$ when adding $i$ above $w$}
22: \textbf{procedure} \textsc{RecursiveUpdate}(w)
23: \hspace*{1em} $(u, v, e) := (\text{the left child of } w, \text{the right child of } w, \text{the parent branch of } w)$
24: \hspace*{1em} for $(X, Y, Z) \in \{(A, B, C), (B, C, A), (C, A, B)\}$ do
25: \hspace*{2em} $Q \leftarrow Q - w_{X\cdot Y\cdot Z}Y€$
26: \hspace*{2em} $w_{X\cdot Y\cdot Z}Y€ \leftarrow w_{X\cdot Y\cdot Z}Y€ + u_Xv_YZ + u_Xv_YY + u_Xv_YZv_Y + u_Xv_YZv_YZ + v_XYv_YZ$
27: \hspace*{2em} $+ (u_Xv_YY + u_Yv_YY - u_Xv_YZ - u_Xv_YZv_Y)$
28: \hspace*{2em} $Q \leftarrow Q + w_{X\cdot Y\cdot Z}Y€$
29: \hspace*{2em} if $w$ is not the root then
30: \hspace*{3em} $(u_Xv_YY, w_Z) \leftarrow ((u_X + v_X)e^{-l(e)}, (u_Y + v_Y)e^{-l(e)}, (u_Z + v_Z)e^{-l(e)})$
31: \hspace*{3em} $w_{X\cdot Y\cdot Z}Y€ \leftarrow u_Xv_YZ + v_XY + u_Xv_YY + u_Yv_YZ + u_Zv_ZY$
32: \hspace*{3em} $w_{Y\cdot Z\cdot X}Y€ \leftarrow (u_Yv_YZ + u_Yv_YZ + u_Zv_ZY)(1 - s(e))$
33: \hspace*{3em} $w_{X\cdot Z\cdot Y}Y€ \leftarrow u_Yv_YZ + u_Yv_YZ + u_Zv_ZY$
34: \hspace*{3em} $w_{X\cdot Y\cdot Z}Y€ \leftarrow u_Xv_YZ + u_Xv_YZ + u_Zv_ZY\left(1 - s(e)\right)$
35: \hspace*{3em} $w_{Y\cdot Z\cdot X}Y€ \leftarrow u_Yv_YZ + u_Yv_YZ + u_Zv_ZY\left(1 - s(e)\right)$
36: \hspace*{3em} $w_{X\cdot Y\cdot Z}Y€ \leftarrow u_Xv_YZ + u_Xv_YZ + u_Zv_ZY\left(1 - s(e)\right)$
37: \hspace*{3em} $w_{X\cdot Y\cdot Z}Y€ \leftarrow u_Xv_YZ + u_Xv_YZ + u_Zv_ZY\left(1 - s(e)\right)$
38: \hspace*{3em} $w_{X\cdot Y\cdot Z}Y€ \leftarrow u_Xv_YZ + u_Xv_YZ + u_Zv_ZY\left(1 - s(e)\right)$
39: \hspace*{3em} $w_{X\cdot Y\cdot Z}Y€ \leftarrow u_Xv_YZ + u_Xv_YZ + u_Zv_ZY\left(1 - s(e)\right)$
40: \hspace*{3em} \textsc{RecursiveUpdate}(the parent of $w$)
41: \textbf{procedure} \textsc{UpdateCounters}(w, X) \Comment*[r]{w is a leaf, X is a color}
42: \hspace*{1em} $e_\cdot :=$ the parent branch of $w$
43: \hspace*{1em} $(w_A, w_B, w_C) \leftarrow (0, 0, 0)$
44: \hspace*{1em} $w_X \leftarrow e^{-l(e)}$
45: \hspace*{1em} \textsc{RecursiveUpdate}(the parent of $w$)
46: \textbf{return} $Q$
Algorithm S2 The Algorithm S2 of $O(n^2kH\log n)$ running time. At start, the function is called as with $\mathcal{L}, \mathcal{G}, r$ as input.

1: procedure NaivePlacement($T, \mathcal{G}, r$)
2:   $W^* \leftarrow$ empty lookup table from tripartitions to their weights
3:   for $i \in \{1, ..., r\}$ do
4:     shuffle $T$
5:     $S_i \leftarrow$ tree with leaves $T_1, T_2,$ and $T_3$
6:   for $j \in \{4, ..., |T|\}$ do
7:     $W_{i,S_i,e} \leftarrow$ Place($T_j, S_i, \mathcal{G}$)
8:   Add all elements of $W_i$ to $W^*$
9: return OptimalTreeDP($T, T, W^*$)

procedure OptimalTreeDP($P, \mathcal{L}, W$)
11: if DPTree($P$) available then
12:   return DPTree($P$)
13: if $|P| = 1$ then
14:   DPScore($P$) $\leftarrow$ 0
15:   DPTree($P$) $\leftarrow$ Singleton rooted tree with leafset $P$
16: else
17:   $X \leftarrow -\infty$
18:   for $A \in \{A: W[A|P - A|L - P] has been computed\}$ do
19:     $S_1 \leftarrow$ OptimalTreeDP($A, \mathcal{L}, W$)
20:     $S_2 \leftarrow$ OptimalTreeDP($P - A, \mathcal{L}, W$)
21:     if DPScore($A$) + DPScore($P - A$) + $W[A|P - A|L - P] > X$ then
22:       $X \leftarrow$ DPScore($A$) + DPScore($P - A$) + $W[A|P - A|L - P]$
23:       DPTree($P$) $\leftarrow$ merge subtrees $S_1$ and $S_2$ at root
24:   DPScore($P$) $\leftarrow X$
25: return DPTree($P$)
Algorithm S3  The DAC algorithm of $O(n^{1.5+k})$ running time given some assumptions. OptimalTreeDP and NaivePlacement are defined in Algorithm S2, and Place is defined in Algorithm S1. At start, the function is called as with $L, G, r$ as input.

1: procedure TwoStepPlacement($T, G, r$)  
2:     $W^* \leftarrow$ empty lookup table from tripartitions to their weights  
3:     for $i \in \{1, \ldots, r\}$ do  
4:         $T_i \leftarrow$ a subsample of $T$ by removing each element independently with probability $1 - 1/\sqrt{|T|}$  
5:         $S_i \leftarrow$ NaivePlacement($T_i, G, \sqrt{|T|}$)  
6:     for $e \in E_{S_i}$ do  
7:         $C_e \leftarrow$ empty list  
8:         for $j \in T - T_i$ do  
9:             $W, S_j, e \leftarrow$ Place($j, S_i, G$)  
10:            add $T_j$ to $C_e$  
11:         $C_\emptyset \leftarrow$ empty list  
12:         $S'_i \leftarrow S_i$  
13:     for $e \in$ branches of $S_i$ do  
14:         $S_e \leftarrow S_i$  
15:         for $j \in C_e$ do  
16:             $W, S_j, e' \leftarrow$ Place($j, S_e, G$)  
17:             if $e' \in S_i - \{e\}$ then  
18:                 add $j$ to $C_\emptyset$  
19:             else  
20:                 $S_e \leftarrow S_e$  
21:             $S'_i \leftarrow S'_i$  
22:         for $j \in C_\emptyset$ do  
23:             $W, S_j, e' \leftarrow$ Place($j, S'_i, G$)  
24:             if $C_\emptyset = \emptyset$ then  
25:                 $W, S'_i, e' \leftarrow$ Place($\emptyset, S'_i, G$)  
26:             Add all elements of $W_i$ to $W^*$  
27:     return OptimalTreeDP($T, T, W^*$)
Proofs

Weighting by support: Proof of Proposition 1 and Theorem 1

For ease of reference, we reproduce Table 2 from the main paper here:

| \( E[\cdot] | \alpha_{G,Q} \) | \( \delta_G(ab|cd) \) | \( \delta_G(ac|bd) \) |
| --- | --- | --- |
| \( \delta_G(ab|cd) \) | \( \geq \frac{1}{2}(1 + 2\theta_Q)(\alpha_{G,Q} + \frac{1}{4}(1 - \alpha_{G,Q})(1 - \beta_Q)) \) | \( \leq \frac{1}{2}(1 + 2\theta_Q)(\frac{1}{2}(1 - \alpha_{G,Q})(1 + \beta_Q)) \) |
| \( \delta_G(ac|bd) \) | \( \geq \frac{1}{2}(1 - \theta_Q)(\frac{1}{2}(1 - \alpha_{G,Q})(1 - \beta_Q)) \) | \( \leq \frac{1}{2}(1 - \theta_Q)(\alpha_{G,Q} + \frac{1}{4}(1 - \alpha_{G,Q})(1 + \beta_Q)) \) |
| \( \delta_G(ad|bc) \) | \( \geq \frac{1}{2}(1 - \theta_Q)(\frac{1}{2}(1 - \alpha_{G,Q})(1 - \beta_Q)) \) | \( \leq \frac{1}{2}(1 - \theta_Q)(\frac{1}{2}(1 - \alpha_{G,Q})(1 + \beta_Q)) \) |

Recall that the expected value and variance of \( \alpha_{G,Q} \) across genes is denoted by \( \bar{\alpha}_Q \) and \( \sigma^2_Q \).

**Proposition 1.** For each estimated gene tree \( G \), \( E[\delta_G(ab|cd) - \delta_G(ac|bd)] \geq \theta_Q \alpha_{G,Q} - \frac{2}{3}(1 - \alpha_{G,Q})\beta_Q \) and \( E[w_G(ab|cd) - w_G(ac|bd)] \geq \frac{1}{9} \theta_Q (3 + 2\beta_Q)(\alpha_{Q}^2 + \sigma^2_Q) + \frac{4}{9}(1 - \alpha_{G,Q})\beta_Q \).

**Proof.** To prove the Proposition, we start with the following lemma.

**Lemma 1.** For each estimated gene tree \( G \) with a given \( \alpha_{G,Q} \),

\[
E[\delta_G(ab|cd) - \delta_G(ac|bd)|w_G(ac|bd)] \geq \theta_Q \alpha_{G,Q} - \frac{2}{3}(1 - \alpha_{G,Q})\beta_Q
\]

and

\[
E[w_G(ab|cd) - w_G(ac|bd)|\alpha_{G,Q}] \geq \frac{1}{9} (3\alpha_{Q} - 2\beta_Q + 2\alpha_{G,Q}\beta_Q + 6)\theta_Q \alpha_{G,Q} - \frac{4}{9}(1 - \alpha_{G,Q})\beta_Q.
\]

**Proof.** From Table 2, we can compute

\[
E[\delta_G(ab|cd) - \delta_G(ac|bd)|\alpha_{G,Q}]
= E\left[ (\delta_G(ab|cd) - \delta_G(ac|bd)) (\delta_G(ab|cd) + \delta_G(ac|bd) + \delta_G(ad|bc) ) |\alpha_{G,Q} \right]
\geq \frac{1}{3} ((1 + 2\theta_Q)\alpha_{G,Q} + \frac{1}{3}(1 - \alpha_{G,Q})(1 - \beta_Q)) - \frac{1}{3} (1 - \theta_Q)\alpha_{G,Q} + \frac{1}{3}(1 - \alpha_{G,Q})(1 + \beta_Q)
= \theta_Q \alpha_{G,Q} - \frac{2}{3}(1 - \alpha_{G,Q})\beta_Q;
\]

similarly,

\[
E[w_G(ab|cd) - w_G(ac|bd)|\alpha_{G,Q}]
= E\left[ (w_G(ab|cd) - w_G(ac|bd)) (\delta_G(ab|cd) + \delta_G(ac|bd) + \delta_G(ad|bc) ) |\alpha_{G,Q} \right]
\geq \frac{1}{3} (1 + 2\theta_Q)\alpha_{G,Q} (\alpha_{G,Q} + \frac{2}{3}(1 - \alpha_{G,Q})(1 - \beta_Q)) + \frac{1}{3}(1 - \alpha_{G,Q})(1 - \beta_Q) \]
\[\geq \frac{1}{3} (1 - \theta_Q)\alpha_{G,Q} (\alpha_{G,Q} + \frac{2}{3}(1 - \alpha_{G,Q})(1 + \beta_Q)) - \frac{1}{3}(1 - \alpha_{G,Q})(1 + \beta_Q) \]  
\[\geq \frac{1}{9} (3\alpha_{Q} - 2\beta_Q + 2\alpha_{G,Q}\beta_Q + 6)\theta_Q \alpha_{G,Q} - \frac{4}{9}(1 - \alpha_{G,Q})\beta_Q.\]

□
From this lemma, we can prove the proposition. First, assume $\alpha_{G,Q}$ is drawn from a discrete distribution. Then,
$$\mathbb{E}[\delta_G(ab|cd) - \delta_G(ac|bd)] = \sum_{\alpha_{G,Q}} \mathbb{E}[\delta_G(ab|cd) - \delta_G(ac|bd)|\alpha_{G,Q}] P(\alpha_{G,Q})$$
$$\geq \sum_{\alpha_{G,Q}} (\theta_2\alpha_{G,Q} - \frac{2}{3}(1 - \alpha_{G,Q})\beta_Q) P(\alpha_{G,Q}) = \theta_2\alpha_Q - \frac{2}{3}(1 - \alpha_Q)\beta_Q$$
and
$$\mathbb{E}[w_G(ab|cd) - w_G(ac|bd)] = \sum_{\alpha_{G,Q}} \mathbb{E}[w_G(ab|cd) - w_G(ac|bd)|\alpha_{G,Q}] P(\alpha_{G,Q})$$
$$\geq \sum_{\alpha_{G,Q}} \left(\frac{1}{9}(3\alpha_{G,Q} - 2\beta_Q + 2\alpha_{G,Q}\beta_Q + 6)\theta_2\alpha_{G,Q} - \frac{4}{9}(1 - \alpha_{G,Q})\beta_Q\right) P(\alpha_{G,Q})$$
$$= \frac{1}{9}\theta_2(3 + 2\beta_Q)\mathbb{E}[\alpha_{G,Q}^2] + \frac{2}{9}(3 - \beta_Q)\theta_2\alpha_Q - \frac{4}{9}(1 - \alpha_Q)\beta_Q$$
$$= \frac{1}{9}\theta_2(3 + 2\beta_Q)(\alpha_Q^2 + \sigma_Q^2) + \frac{2}{9}(3 - \beta_Q)\theta_2\alpha_Q - \frac{4}{9}(1 - \alpha_Q)\beta_Q.$$ 

It is straightforward to change these calculations to use integral instead of sum and $P(\alpha_{G,Q})$ to the PDF in the case that the distribution of $\alpha_{G,Q}$ is continuous. 

**Theorem 1.** Given estimated gene trees furnished with support generated under MSC+Error+Support model, there exist conditions where (3) guarantee a statistically consistent estimator of $S^*$ but (2) does not, and the reverse is not true.

**Proof.** Recall that (1) states
$$W(S,G) := \sum_{G \subseteq [Q \cup \xi(S)]} \sum_{Q \subseteq \xi(S)} w_G(S \mid Q).$$

It means that in order to produce a statistically consistent estimator using (1), the following equation must be satisfied for the true species tree topology $S^*$ and any species tree topology $S$:
$$\mathbb{E}[W(S^*,G) - W(S,G)] = |G| \sum_{Q \subseteq \xi(S)} \mathbb{E}[w_G(S^* \mid Q) - w_G(S \mid Q)] \geq 0 \quad (9)$$

Notice that proving for any quartet $Q = \{a,b,c,d\}$ we have $\mathbb{E}[w_G(ab|cd) - w_G(ac|bd)] \geq 0$ and $\mathbb{E}[w_G(ab|cd) - w_G(ac|bd)] \geq 0$ where $S^* \mid Q = ab|cd$ is sufficient to prove (9); on the other hand, proving for any quartet $Q = \{a,b,c,d\}$ where the internal branch of $S^* \mid Q$ corresponds to only one branch in $S^*$, we have $\mathbb{E}[w_G(ab|cd) - w_G(ac|bd)] \geq 0$ and $\mathbb{E}[w_G(ab|cd) - w_G(ac|bd)] \geq 0$ where $S^* \mid Q = ab|cd$ is necessary to prove (9).

Thus, from Proposition 1, we have guaranteed statistical consistency for wASTRAL for support under
$$D = \cap_{Q \subseteq \xi(S)} \{ (\theta_Q, \alpha_Q, \sigma_Q, \beta_Q) \in [0,1]^4 : \frac{1}{9}\theta_Q(3 + 2\beta_Q)(\alpha_Q^2 + \sigma_Q^2) + \frac{2}{9}(3 - \beta_Q)\theta_2\alpha_Q - \frac{4}{9}(1 - \alpha_Q)\beta_Q \geq 0 \}.$$

Similarly, we have guaranteed statistical consistency for unweighted ASTRAL under
$$D' = \cap_{Q \subseteq \xi(S)} \{ (\theta_Q, \alpha_Q, \sigma_Q, \beta_Q) \in [0,1]^4 : \theta_2\alpha_Q - \frac{2}{3}(1 - \alpha_Q)\beta_Q \geq 0 \}.$$ 

To prove Theorem 1, we only need to prove that $D'$ is a proper subset of $D$. 

□
We can prove $D' \subseteq D$, as for any $Q$, if $(\theta_Q, \bar{\alpha}_Q, \sigma, \beta_Q) \in [0,1]^4$ and $\theta_Q \bar{\alpha}_Q - \frac{2}{3} (1 - \bar{\alpha}_Q) \beta_Q \geq 0$, then
\[
\frac{1}{9} \theta_Q (3 + 2 \beta_Q)(\bar{\alpha}_Q^2 + \sigma^2) + \frac{2}{9} (3 - \beta_Q) \theta_Q \bar{\alpha}_Q - \frac{4}{9} (1 - \bar{\alpha}_Q) \beta_Q \\
= \frac{1}{9} \theta_Q (3 + 2 \beta_Q)(\bar{\alpha}_Q^2 + \sigma^2) + \frac{1}{3} \theta_Q (1 - \theta_Q) \bar{\alpha}_Q^2 + (\frac{1}{3} \theta_Q \bar{\alpha}_Q + \frac{2}{3}) (\theta_Q \bar{\alpha}_Q - \frac{2}{3} (1 - \bar{\alpha}_Q) \beta_Q) \geq 0.
\]
We can also prove $D' \neq D$, as if for some $Q$, $\theta_Q = 0.25, \bar{\alpha}_Q = 0.5, \beta_Q = 0.4$, 
\[
\theta_Q \bar{\alpha}_Q - \frac{2}{3} (1 - \bar{\alpha}_Q) \beta_Q = -\frac{1}{120} < 0
\]
and
\[
\frac{1}{9} \theta_Q (3 + 2 \beta_Q)(\bar{\alpha}_Q^2 + \sigma^2) + \frac{2}{9} (3 - \beta_Q) \theta_Q \bar{\alpha}_Q - \frac{4}{9} (1 - \bar{\alpha}_Q) \beta_Q = \frac{7}{720} + \frac{19}{180} \sigma^2 > 0.
\]
Thus $D'$ is a proper subset of $D$ and we conclude the proof. □
Weighting by length: Proof of Propositions 2 and 3 and Theorem 2

Before providing the proofs, we remind the reader of one property of the coalescent model. According to the coalescent model, at any point along a branch of the species tree with i gene tree lineages, the time (i.e., distance) \( x \) to the next coalescent event, reducing the number of lineages to \( i - 1 \), is exponentially distributed with the rate \( \binom{i}{2} \), resulting in probability density function (PDF):

\[
\frac{i(i-1)}{2} e^{-\frac{x+1}{x}},
\]

and the two lineages that coalesce are independent of \( x \).

**Proposition 2.** For a true quartet species tree \( S^* \) with topology \( ab|cd \) and input gene trees \( G \) generated under the naive model with any multiplier \( \lambda \), let \( f \) be the distance between anchors of \( S^* \). As \( f \to 0 \), given \( k=\Theta(f^{-2}) \) gene trees, we have \( \text{Var}[X_G]=\Theta(f) \) and

\[
\frac{\text{E}[X_G]}{\sqrt{\text{Var}[X_G]}} = \frac{1+4\lambda}{1+2\lambda} \sqrt{\frac{3}{2}} f + O(f^2).
\]

**Proof.** We analyze balanced and unbalanced trees separately.

Case 1: Unbalanced trees (i.e., the root of \( S^* \) has a terminal branch as a child). W.o.l.g., we assume the root branch is located on branch leading to \( d \).

Let \( p,q \), and \( r \) be the MRCA nodes of \( (a,b) \), \( (a,c) \), and \( (a,d) \) on rooted species tree \( S^* \), respectively. Let \( p' \) and \( r' \) be the points of coalescence of leaves \( a,b \) and leaves \( c,d \) on the rooted gene tree \( G \), respectively. Let \( x, y_0 \), and \( z \) be the CU difference in heights of points \( (p,p') \), \( (q,r) \), and \( (r,r') \), respectively. Note that \( f \) is the length of \( (p,q) \). Let \( L:=L_S(a,p)+L_S(b,p)+L_S(c,r)+L_S(d,r) \). Notice that \( l_G(a,p)+l_G(b,p)+l_G(c,r)+l_G(d,r)=\lambda L \) and \( l_G(a,b)+l_G(c,d)=\lambda(2x+2z+L) \).

Let \( f_X(x) \) be the probability density that \( x \) is the CU difference in heights of \( (p,p') \) and \( p' \) is the lowest point of coalescence. Notice that by (10):

\[
f_X(x) = \begin{cases} 
  e^{-x} & 0 \leq x \leq f \\
  \frac{1}{f} \left( e^{-f} \binom{2}{3} e^{-\left(\binom{3}{3}(x-f)\right)} \right) = e^{-3x+2f} & f \leq x \leq f + y_0 \\
  \frac{1}{f} \left( e^{-f} e^{-\left(\binom{3}{3}y_0 \binom{2}{3} e^{-\left(\binom{3}{3}(x-f-y_0)\right)} \right)} = e^{-6x+5f+3y_0} & f + y_0 \leq x 
\end{cases}
\]

Let \( f_{Z|X}(z;x) \) be the probability density that \( z \) is the CU difference in heights of \( (r,r') \), conditioned on that \( x \) is the CU difference in heights of \( (p,p') \) and \( p' \) is the lowest point of coalescence. Notice that:

\[
f_{Z|X}(z;x) = \begin{cases} 
  e^{-z} & 0 \leq x \leq f + y_0 \text{ and } 0 \leq z \\
  e^{-z} \left( e^{-z-(x-f-y_0)} \right) = e^{-z + x - f - y_0} & 0 \leq x - f - y_0 \leq z
\end{cases}
\]

We specify three coalescence scenarios by indicator functions \( \delta_1, \delta_2, \delta_3 \): i) \( \delta_1 \) indicates \( 0 \leq x < f \); ii) \( \delta_2 \) indicates \( f \leq x < f + y_0 \); iii) \( \delta_3 \) indicates \( f + y_0 \leq x \).

Note that

\[
\text{E}[w_G(ab|cd)] = \text{E}[(\delta_1 + \delta_2 + \delta_3)w_G(ab|cd)]
\]

\[
\text{E}[w_G^2(ab|cd)] = \text{E}[(\delta_1 + \delta_2 + \delta_3)w_G^2(ab|cd)].
\]
FIG. S23. Illustration of the unbalanced case. Lengths in CU/SU units are denoted in blue/red. Branches in green have a
total length $L/\lambda L$ in CU/SU units. The right-hand side shows the position of $p'$ in relation to $q$ and $r$ in various cases.

Similarly, since only scenarios 2 and 3 have deep coalescence events that may lead to gene tree
disagreement with the species tree, and by the symmetry of all three topologies under scenarios 2 and 3,

$$E[w_G(ac|bd)] = E[(\delta_2 + \delta_3)u_G(ab|cd)]$$
$$E[w_G^2(ac|bd)] = E[(\delta_2 + \delta_3)u_G^2(ab|cd)].$$

Thus,

$$E[X_G] = E[w_G(ab|cd)] - E[w_G(ac|bd)] = E[\delta_1 w_G(ab|cd)], \quad (11)$$

and since $w_G(ab|cd)w_G(ac|bd) = 0$,

$$\text{Var}[X_G] = E[X_G^2] - E^2[X_G] = E[w_G^2(ab|cd)] + E[w_G^2(ac|bd)] - E^2[X_G]$$
$$= E[(\delta_1 + 2\delta_2 + 2\delta_3)w_G^2(ab|cd)] - E^2[X_G]. \quad (12)$$

We next compute both elements of (11) as well as some elements of (12) (others will not be necessary).

- $\delta_1$: When $G$ has topology $ab|cd$, $p'$ must be the lowest point of coalescence. Thus,

$$E[\delta_1 w_G(ab|cd)]$$
$$= \int_0^f \int_0^{\infty} e^{-\lambda(2x+2z+L)} f_X(x) f_Z(x) dz dx$$
$$= \int_0^f \int_0^{\infty} e^{-\lambda(2x+2z+L)} e^{-x} e^{-z} dz dx$$
$$= e^{-\lambda f(1-e^{-2\lambda})} f \left(\frac{1+2\lambda}{\lambda^2}\right).$$
\[ \mathbb{E}[\delta_1 w_G(ab|cd)] \leq \mathbb{E}[\delta_1 w_G(ab|cd)] = O(f). \]

\textbullet \ \delta_2: \text{When} \ G \text{ has topology } ab|cd, \ p' \text{ must be the lowest point of coalescence. Thus,}
\[
\mathbb{E}[\delta_2 w_G(ab|cd)] = \int_{f+y_0}^{+\infty} e^{-\lambda(x+4x+2L)} f_X(x) f_G(z) dz dx \\
= \int_{f+y_0}^{+\infty} e^{-\lambda(x+4x+2L)} e^{-3x+2f} e^{-z} dz dx \\
= \int_{f+y_0}^{+\infty} e^{-\lambda(x+4x+2L)} e^{-3x+2f} e^{-z} dz dx \\
= \frac{1}{(1+4\lambda)(1+4\lambda)} e^{-(1+4\lambda)f-2\lambda L}.
\]

\textbullet \ \delta_3: \text{When} \ G \text{ has the topology } ab|cd, \text{ either } p' \text{ or } q' \text{ must be the lowest point of coalescence, and by symmetry, the two cases must have the same PDFs. Thus,}
\[
\mathbb{E}[\delta_3 w_G(ab|cd)] = \int_{f+y_0}^{+\infty} e^{-\lambda(x+4x+2L)} f_X(x) f_G(z) dz dx \\
= \int_{f+y_0}^{+\infty} e^{-\lambda(x+4x+2L)} e^{-3x+2f} e^{-z} dz dx \\
= \int_{f+y_0}^{+\infty} e^{-\lambda(x+4x+2L)} e^{-3x+2f} e^{-z} dz dx \\
= \frac{1}{(3+4\lambda)(1+4\lambda)} e^{-(1+4\lambda)f-2\lambda L}.
\]

Replacing in (11), we get
\[
\mathbb{E}[X_G] = \mathbb{E}[\delta_1 w_G(ab|cd)] = e^{-\lambda L (1-\delta^2)} \frac{1}{(1+2\lambda)^2} = e^{-\lambda L} \frac{1}{1+2\lambda} f + O(f^2);
\]
and replacing in (12), we get
\[
\text{Var}[X_G] = \mathbb{E}[\delta_1 + 2\delta_2 + 2\delta_3 w_G(ab|cd)] - \mathbb{E}[X_G] = \mathbb{E}[2\delta_2 + \delta_3 w_G(ab|cd)] + O(f) \\
= 2e^{-\lambda L} + O(f) = \frac{2e^{-\lambda L}}{(3+4\lambda)(1+4\lambda)} + O(f).
\]
from which our assumption of Var[\(X_G\)] = \(\Omega(1)\) follows.

Case 2: Balanced tree.

Let \(p,q\), and \(r\) be the MRCA nodes of \((a,b)\), \((c,d)\), and \((a,d)\) on rooted species tree \(S^*\), respectively. Let \(p'\) and \(q'\) be the points of coalescence of leaves \(a,b\) and leaves \(c,d\) on the rooted gene tree \(G\), respectively. Let \(x, y, z, y_0\) be the CU difference in heights of points \((p,p')\), \((p,r)\), \((q,q')\), and \((q,r)\), respectively. Note that \(f = x + y\) is CU length of path \((p,q)\). Let \(L := l_S(a,p) + l_S(b,p) + l_S(c,q) + l_S(d,q)\). Notice that \(l_G(a,p) + l_G(b,p) + l_G(c,q) + l_G(d,q) = \lambda L\) and \(l_G(a,b) + l_G(c,d) = \lambda (2x + 2y + L)\).

We specify three coalescence scenarios by indicator functions \(\delta_1, \delta_2, \delta_3\): i) \(\delta_1\) indicates \(0 \leq x < x_0\); ii) \(\delta_2\) indicates \(x_0 \leq x, 0 \leq y < y_0\); iii) \(\delta_3\) indicates \(x_0 \leq x, y_0 \leq y\).

Note that
\[
\mathbb{E}[w_G(ab|cd)] = \mathbb{E}[\delta_1 + \delta_2 + \delta_3 w_G(ab|cd)] \\
\mathbb{E}[w^2_G(ab|cd)] = \mathbb{E}[\delta_1^2 + \delta_2^2 + \delta_3 w^2_G(ab|cd)].
\]
Similarly, since only scenarios 3 have deep coalescence events that may lead to gene tree disagreement with the species tree, and by the symmetry of all three topologies under scenarios 3,

$$E\left[w_{G}(ac|bd)\right] = E\left[\delta_{1}w_{G}(ab|cd)\right]$$

Thus,

$$E[X_{C}] = E\left[w_{G}(ab|cd)\right] - E\left[w_{G}(ac|bd)\right] = E\left[(\delta_{1} + \delta_{2})w_{G}(ab|cd)\right]; \quad (13)$$

and since $w_{G}(ab|cd)w_{G}(ac|bd) = 0$,

$$\text{Var}[X_{C}] = E[X_{C}^{2}] - E^{2}[X_{C}] = E\left[w_{G}^{2}(ab|cd) + w_{G}^{2}(ac|bd)\right] - E^{2}[X_{C}]$$

$$= E\left[(\delta_{1} + \delta_{2} + 2\delta_{3})w_{G}^{2}(ab|cd)\right] - E^{2}[X_{C}]. \quad (14)$$

- $\delta_{1}$: Here,

$$E[\delta_{1}w_{G}(ab|cd)] = \int_{0}^{x_{0}} \int_{0}^{\infty} e^{-\lambda(2x+2y+L)} e^{-\lambda x} e^{-\lambda y} \, dx \, dy$$

$$= e^{-\lambda L}(1 - e^{-(1+2\lambda)x_{0}}) = e^{-\lambda Lx_{0}} + O(x_{0}^{2}) = e^{-\lambda Lx_{0}} + O(f^{2});$$

and

$$E[\delta_{1}w_{G}^{2}(ab|cd)] \leq E[\delta_{1}w_{G}(ab|cd)] = O(f).$$
\[ \delta_2 \text{: Here,} \]
\[ E[\delta_2 w_G(abc)] = \int_0^{+\infty} \int_0^{y_0} e^{-\lambda (2x + 2y + L)} e^{-x} e^{-y} dy dx \]
\[ = e^{-\lambda L} (1 - e^{-(1+2\lambda) y_0}) e^{-(1+2\lambda) x_0} \frac{(1+2\lambda)^2}{(1+2\lambda)} = e^{-\lambda L} y_0 = (1+2\lambda) + O(f^2); \]

and
\[ E[\delta_2 w_G^2(abc)] \leq E[\delta_2 w_G(abc)] = O(1). \]

\[ \delta_3 \text{: Similar to the unbalanced case, when } G \text{ has the topology } ab|cd, \text{ either } p' \text{ or } q' \text{ must be the lowest point of coalescence, and by symmetry, the two cases must have the same PDFs. Thus,} \]
\[ E[\delta_3 w_G^2(ab|cd)] = \int_{x_0}^{+\infty} \int_{x_0+y_0}^{+\infty} e^{-\lambda (2x_0 + 2y_0 + 2L)} e^{-x_0} e^{-y_0} e^{-x-x_0+y} e^{-y-y_0+y_0} dy dx \]
\[ = \int_{x_0}^{+\infty} e^{-4\lambda (x-x_0+y_0)} e^{-(1+2\lambda) x_0} e^{-y_0} e^{-y+x-x_0+y_0} \frac{1}{1+2\lambda} dy dx \]
\[ = \frac{1}{(3+4\lambda)(1+2\lambda)} e^{-(1+4\lambda)(x_0+y_0)-2\lambda} = \frac{1}{(3+4\lambda)(1+2\lambda)} e^{-(1+4\lambda)f-2\lambda}. \]

Replacing in (13), we get
\[ E[X_G] = E[(\delta_1 + \delta_2) w_G(ab|cd)] = \frac{e^{-\lambda L}(x_0+y_0)}{1+2\lambda} + O(f^2) = \frac{e^{-\lambda L} f}{1+2\lambda} + O(f^2); \]

and replacing in (14), we get
\[ \text{Var}[X_G] = E[(\delta_1 + \delta_2 + 2\delta_3) w_G^2(ab|cd)] - E^2[X_G] \]
\[ = E[2\delta_3 w_G^2(ab|cd)] = O(f) \]
\[ = 2e^{-(1+4\lambda)f-2\lambda}\frac{2\lambda}{(3+4\lambda)(1+4\lambda)} + O(1). \]

from which our assumption of \( \text{Var}[X_G] = \Theta_f(1) \) follows.

Thus, in both balanced and unbalanced cases,
\[ \frac{E[X_G]}{\sqrt{\text{Var}[X_G]}} = \frac{\frac{e^{-\lambda L}}{1+2\lambda} f + O(f^2)}{\sqrt{2\frac{e^{-2\lambda L}}{(1+2\lambda)(3+4\lambda)} + O(1)}} = \sqrt{\frac{1+4\lambda + 4\lambda^2}{3(1+2\lambda)^2}} \sqrt{\frac{3}{2} f + O(f^2)} \]

\[ \square \]

**Proposition 3.** For a true quartet species tree \( S^* \) with topology \( ab|cd \) and input gene trees \( G \) generated under the variable rate model, let \( f \) be the distance between anchors of \( S^* \) and \( L \) be the total length of all other branches. Assume that for every branch segment \( I \), the variance of its multiplier is bounded above: \( \text{Var}(\Lambda_{S'}^I) \leq \varepsilon^2 \) where \( \varepsilon^2 = \frac{e^{-\lambda L}}{(1+32\lambda^2)+(6+32\lambda+32\lambda^2)^2} \left( \frac{20(\lambda+\lambda^2)}{(1+2\lambda)^2} \right) \). As \( f \to 0 \), given \( k = \Theta(f^{-2}) \) gene trees, we have \( \text{Var}[X_G] = \Theta_f(1) \) and
\[ \frac{E[X_G]}{\sqrt{\text{Var}[X_G]}} \geq \sqrt{\frac{\pi}{2}} (1 - \frac{4\lambda^2}{(1+4\lambda)^2})^{-\frac{1}{2}} f + O(f^2). \]

**Proof.** We follow the same logic in proof of Proposition 2.
Case 1: Unbalanced trees. Let $P(x)$ be functions to random variables denoting SU difference in heights of points $(p, p')$ where $p'$ is $x$ CU distance above $p$; let $R(z)$ be functions to random variables denoting SU difference in heights of points $(r, r')$ where $r'$ is $z$ CU distance above $r$. Note that $P(f + y_0) + R(z) = P(f + y_0 + z)$ where $P(f + y_0)$ denote the SU length of $(p, r)$. Let random variable $\Lambda := l_{S^\uparrow}(a, p) + l_{S^\uparrow}(b, p) + l_{S^\uparrow}(c, r) + l_{S^\uparrow}(d, r)$ be the total SU terminal branch lengths and the constant value $L$ be the CU distance corresponding to $\Lambda$.

**FIG. S25.** Illustration of the unbalanced case. Lengths in CU/SU units are denoted in blue/red. Branches in green have a total length $L/\Lambda$ in CU/SU units. The right-hand side shows the position of $p'$ in relation to $q$ and $r$ in various cases.

- $\delta_1$: When $G$ has topology $ab|cd$, $p'$ must be the lowest point of coalescence. Thus,

\[
\mathbb{E}[\delta_1 w_G(ab|cd)] = \mathbb{E}\left[\int_0^{+\infty} \int_0^{+\infty} e^{-2P(x)-2R(z)-\Lambda} f_X(x) f_Z|X(z;x) dx dz\right]
\]

\[
= \mathbb{E}\left[\int_0^{+\infty} \int_0^{+\infty} e^{-2P(x)-2R(z)-\Lambda} e^{-x-z} dx dz\right]
\]

\[
= \mathbb{E}\left[\int_0^{+\infty} \int_0^{+\infty} e^{-2P(x)-2R(z)-\Lambda-x-z} dx dz\right];
\]

and

\[
\mathbb{E}[\delta_1 w_G^2(ab|cd)] \leq \mathbb{E}[\delta_1 w_G(ab|cd)] = O(f).
\]
\[ \delta_2: \text{When } G \text{ has topology } ab|cd, \text{ either } p' \text{ or } q' \text{ must be the lowest point of coalescence, and by symmetry, the two cases must have the same PDFs. Thus,} \\
\mathbb{E}[\delta_2w_2^2(ab|cd)] \\
= \mathbb{E}\left[ \int_f^{f+y_0} \int_0^{+\infty} e^{-4P(x)-4R(z)z} f_X(x)f_{Z|X}(z;x) \, dz \, dx \right] \\
= \mathbb{E}\left[ \int_f^{f+y_0} \int_0^{+\infty} e^{-4P(x)-4R(z)z} e^{-3x+2f} \, dz \, dx \right] \\
= \int_f^{f+y_0} \int_0^{+\infty} \mathbb{E}[e^{-4P(x)-4R(z)z}e^{-3x+2f}] \, dz \, dx. \]

Replacing in (11), by Jensen’s inequality, we get

\[ \mathbb{E}[X_G] = \mathbb{E}[\delta_1w_G(ab|cd)] = \mathbb{E}\left[ \int_f^{f+y_0} \int_0^{+\infty} e^{-2P(x)-2R(z)z} \, dz \, dx \right] \]

\[ \geq \int_0^{f+y_0} \int_0^{+\infty} \mathbb{E}[e^{-2P(x)-2R(z)z}] \, dx \, dz \]

\[ = \int_0^{f+y_0} \int_0^{+\infty} e^{-2x-2z} \, dx \, dz \]

\[ = e^{-\lambda L(1-e^{-(1+2\lambda)f})} / (1+2\lambda)^2 = e^{-\lambda L} + O(f^2). \]

And replacing in (12), we get

\[ \text{Var}[X_G] = \mathbb{E}[\delta_1 + 2\delta_2 + 2\delta_3]w_2^2(ab|cd) - \mathbb{E}^2[X_G] = \mathbb{E}[2(\delta_2 + \delta_3)w_2^2(ab|cd)] + O(f) \]

\[ = \int_f^{f+y_0} \int_0^{+\infty} \mathbb{E}[e^{-4P(x)-4R(z)z}e^{-3x+2f}] \, dz \, dx \]

\[ + \int_f^{f+y_0} \int_0^{+\infty} \mathbb{E}[e^{-4P(x)-4R(z)z}e^{-3x+2f}] \, dz \, dx + O(f), \]

from which our assumption of \[\text{Var}[X_G] = \Theta_f(1)\] follows.

Let \(F_P(u;x), F_R(v;z), \text{ and } F_\Lambda(w)\) be the CDF of \(P(x), \text{ and } R(z), \text{ and } \Lambda\) respectively; let \(F_{PRA}(u,v,w;x,z)\) and \(F_{PRA}(u,v,w;x,z)\) be the joint CDF and the joint PDF. Let \(F_1^{-1}(t;x), F_R^{-1}(t;z), \text{ and } F_\Lambda^{-1}(t)\) be the inverse function of CDF of \(P(x), R(z), \text{ and } \Lambda\).
Then,

\[
E[e^{-2(2P(x)+2R(z)+\Lambda)}] \\
= \int_0^{+\infty} \int_0^{+\infty} \int_0^{+\infty} e^{-2(2u+2v+w)} F_{P_{R \Lambda}}(u,v,w;x,z) dw dv du \\
= \int_0^{+\infty} \int_0^{+\infty} \int_0^{+\infty} e^{-2(2u+2v+w)} \frac{\partial^3 F_{P_{R \Lambda}}}{\partial u \partial v \partial w} dw dv du \\
= \int_0^{+\infty} \int_0^{+\infty} \left( e^{-2(2u+2v+w)} \frac{\partial^3 F_{P_{R \Lambda}}}{\partial u \partial v \partial w} \right) dw dv du \\
+ \int_0^{+\infty} \int_0^{+\infty} \frac{\partial e^{-2(2u+2v+w)} \frac{\partial^3 F_{P_{R \Lambda}}}{\partial u \partial v \partial w}}{\partial u} dw dv du \\
= \int_0^{+\infty} \int_0^{+\infty} \left( e^{-2(2u+2v+w)} \frac{\partial^3 F_{P_{R \Lambda}}}{\partial u \partial v \partial w} \right) dw dv du \\
+ \int_0^{+\infty} \int_0^{+\infty} \frac{\partial e^{-2(2u+2v+w)} \frac{\partial^3 F_{P_{R \Lambda}}}{\partial u \partial v \partial w}}{\partial u} dw dv du \\
= \int_0^{+\infty} \int_0^{+\infty} \left( e^{-2(2u+2v+w)} \frac{\partial^3 F_{P_{R \Lambda}}}{\partial u \partial v \partial w} \right) dw dv du \\
+ \int_0^{+\infty} \int_0^{+\infty} \frac{\partial e^{-2(2u+2v+w)} \frac{\partial^3 F_{P_{R \Lambda}}}{\partial u \partial v \partial w}}{\partial u} dw dv du \\
= \int_0^{+\infty} \int_0^{+\infty} \left( e^{-2(2u+2v+w)} \frac{\partial^3 F_{P_{R \Lambda}}}{\partial u \partial v \partial w} \right) dw dv du \\
+ \int_0^{+\infty} \int_0^{+\infty} \frac{\partial e^{-2(2u+2v+w)} \frac{\partial^3 F_{P_{R \Lambda}}}{\partial u \partial v \partial w}}{\partial u} dw dv du \\
\]

Thus, for any \(0 < t_0 < 1\),

\[
E[e^{-2(2P(x)+2R(z)+\Lambda)}] \\
\leq \int_0^{1} e^{-2(2P_{R^{-1}(t,x)}+2R_{R^{-1}(t,z)}+F_{R^{-1}}(t))} dt \\
\leq \int_0^{t_0} e^{-2(2P_{R^{-1}(t_0,x)}+2R_{R^{-1}(t_0,z)}+F_{R^{-1}}(t_0))} dt + \int_1^{+\infty} e^{-2(2P_{R^{-1}(t_0,x)}+2R_{R^{-1}(t_0,z)}+F_{R^{-1}}(t_0))} dt \\
\leq t_0 e^{-2(2P_{R^{-1}(t_0,x)}+2R_{R^{-1}(t_0,z)}+F_{R^{-1}}(t_0))}.
\]

By Chebyshev’s inequality (using \(t_0^{1/2}\) as the constant), \(F_{R^{-1}}(t_0;x) \geq (\lambda - \frac{1}{\sqrt{m}})x\), \(F_{R^{-1}}(t_0;z) \geq (\lambda - \frac{1}{\sqrt{m}})y\), and \(F_{R^{-1}}(t_0) \geq (\lambda - \frac{1}{\sqrt{m}})L\). Thus,

\[
E[e^{-2(2P(x)+2R(z)+\Lambda)}] \leq t_0 + e^{-\lambda + \frac{1}{\sqrt{m}}(4x+4z+2L)}.
\]
Thus,
\[
\text{Var}[X_G] \leq \int_{f'}^{g_f} \int_{0}^{f+y_0} \left( t_0 + e^{-(\lambda + 1/\sqrt{\alpha f} + 2f d + 2)} \right) dy dx
\]
\[+ \int_{f'}^{g_f} \int_{x-f-y_0}^{x} \left( t_0 + e^{-(\lambda + 1/\sqrt{\alpha f} + 2f d + 2)} \right) dy dx + O(f)
\]
\[= \int_{f'}^{g_f} \left( 2t_0 e^{-f} + 2f \right) dx + \int_{f'}^{g_f} \left( 2t_0 e^{-f} + 2f \right) dx + O(f)
\]
\[= \frac{2}{3} t_0 (e^{-f} - e^{-3f}) + \frac{2}{1 + 4\lambda - d^2 \sqrt{\alpha f}} (1 + 2f \lambda - d^2 \sqrt{\alpha f}) f
\]
\[+ \frac{2}{3} t_0 (e^{-f} - e^{-3f}) + \frac{2}{1 + 4\lambda - d^2 \sqrt{\alpha f}} (1 + 2f \lambda - d^2 \sqrt{\alpha f}) f
\]
\[= \frac{2}{3} t_0 (e^{-f} - e^{-3f}) + \frac{2}{1 + 4\lambda - d^2 \sqrt{\alpha f}} (1 + 2f \lambda - d^2 \sqrt{\alpha f}) f
\]
\[+ \frac{2}{3} t_0 (e^{-f} - e^{-3f}) + \frac{2}{1 + 4\lambda - d^2 \sqrt{\alpha f}} (1 + 2f \lambda - d^2 \sqrt{\alpha f}) f
\]
\[= \frac{2}{3} t_0 (e^{-f} - e^{-3f}) + \frac{2}{1 + 4\lambda - d^2 \sqrt{\alpha f}} (1 + 2f \lambda - d^2 \sqrt{\alpha f}) f
\]
\[= \frac{2}{3} t_0 (e^{-f} - e^{-3f}) + \frac{2}{1 + 4\lambda - d^2 \sqrt{\alpha f}} (1 + 2f \lambda - d^2 \sqrt{\alpha f}) f
\]

Case 2: Balanced tree. Let \( P(x) \) be functions to random variables denoting SU difference in heights of points \( (p, p') \) where \( p' \) is \( x \) CU distance above \( p \); let \( Q(y) \) be functions to random variables denoting SU difference in heights of points \( (q, q') \) where \( q' \) is \( y \) CU distance above \( q \). Note that \( P(x_0 + z) - P(x_0) = Q(z) - Q(0) \) where \( P(x_0) \) and \( Q(y_0) \) denote the SU length of \( (p, r) \) and \( (q, r) \), respectively. Let random variable \( \Lambda := \text{SU}(a, p) + \text{SU}(b, p) + \text{SU}(c, q) + \text{SU}(d, q) \) be the total SU terminal branch lengths and the constant value \( L \) be the CU distance corresponding to \( \Lambda \).

- \( \delta_1 \): Here,
\[
E[\delta_1 W_G(ab|cd)] = E\left[ \int_{x_0}^{+\infty} \int_{0}^{f+y_0} e^{-2P(x)-2Q(y)-2\Lambda} e^{-s} e^{-y} dy dx \right];
\]
and
\[
E[\delta_1 W_G^2(ab|cd)] = E[\delta_1 W_G(ab|cd)] = O(f).
\]

- \( \delta_2 \): Here,
\[
E[\delta_2 W_G(ab|cd)] = E\left[ \int_{x_0}^{+\infty} \int_{0}^{f+y_0} e^{-2P(x)-2Q(y)-2\Lambda} e^{-s} e^{-y} dy dx \right];
\]
and
\[
E[\delta_2 W_G^2(ab|cd)] = E[\delta_2 W_G(ab|cd)] = O(f).
\]

- \( \delta_3 \): Similar to the unbalanced case, when \( G \) has the topology \( ab|cd \), either \( p' \) or \( q' \) must be the lowest point of coalescence, and by symmetry, the two cases must have the same PDFs. Thus,
\[
E[\delta_3 W_G^2(ab|cd)] = E\left[ \int_{x_0}^{+\infty} \int_{x_0+y_0}^{+\infty} e^{-4P(x)-4Q(y)-4\Lambda} e^{-6x+6x_0} e^{-y} dy dx \right]
\]
\[= \int_{x_0}^{+\infty} \int_{x_0+y_0}^{+\infty} E\left[ \int_{x_0+y_0}^{+\infty} e^{-4P(x)-4Q(y)-4\Lambda} e^{-5x-y+4\Lambda} dy dx \right].
\]
Replacing in (13), we get

\[ E[X_G] = E[(\delta_l + \delta_u)w_{G|ab|cd}] = E\left[ \int_0^{x_0} \int_0^{y_0} e^{-2P(x)} - 2Q(y) - \Lambda e^{-x}e^{-y} dy dx + \int_{x_0}^{+\infty} \int_0^{y_0} e^{-2P(x)} - 2Q(y) - \Lambda e^{-x}e^{-y} dy dx \right] \]

\[ \geq \int_0^{x_0} \int_0^{y_0} e^{-2\lambda x - 2\lambda y - \Lambda L e^{-x}e^{-y}} dy dx + \int_{x_0}^{+\infty} \int_0^{y_0} e^{-2\lambda x - 2\lambda y - \Lambda L e^{-x}e^{-y}} dy dx \]

\[ = \frac{(x_0 + y_0)e^{-\Lambda L}}{1 + 2\lambda} + O(f^2) \approx \frac{\lambda L}{1 + 2\lambda} + O(f^2); \]
and replacing in (14), for any $0 < t_0 < 1$,

\[ \text{Var}[X_G] = \mathbb{E}\left[ (d_1 + 3t_0 + 2 \lambda_0) \mathbb{E}[c] \right] - \mathbb{E}^2[X_G] \]

\[ = \mathbb{E}\left[ 2 \alpha_0 \mathbb{E}[c] \mathbb{E}[d] \right] + O(f) \]

\[ \leq \int_0^\infty \int_0^x \int_{x-x_0+y_0} \mathbb{E} \left[ e^{-4P(x)-4Q(y)-2\lambda_0} \right] \int_{x-x_0+y_0} e^{-6x-y+4z} dx dy dx \]

\[ \leq \int_0^\infty \int_0^x \int_{x-x_0+y_0} \left( t_0 + e^{-\lambda \sqrt{y}} (4z+4y+2L) \right) e^{-6x-y+4z} dx dy dx + O(f) \]

\[ = \int_0^\infty \left( 4e^{-6x-y+5z} + \frac{4}{1+4\lambda-\frac{2\lambda}{\sqrt{y}}} \right) dx + O(f) \]

\[ = \frac{4}{6} e^{-x_0+y_0} + \frac{4}{(1+4\lambda-\frac{2\lambda}{\sqrt{y}})(6+8\lambda-\frac{8\lambda}{\sqrt{y}})} e^{-x_0+y_0} + O(f) \]

\[ \leq \frac{2}{3} t_0 + \frac{2e^{-2\lambda L(\lambda-\frac{\lambda}{\sqrt{y}})}}{(1+4\lambda-\frac{2\lambda}{\sqrt{y}})(3+4\lambda-\frac{4\lambda}{\sqrt{y}})} + O(f) , \]

from which our assumption of $\text{Var}[X_G] = \Theta(f)$ follows. Thus, for both balanced and unbalanced trees, the variance is bounded by the same expression, and thus in both cases,

\[ \text{Var}[X_G] \leq \frac{2}{3} t_0 + \frac{e^{-2\lambda L}}{(1+4\lambda)(3+4\lambda)} + O(f) \]

\[ \leq \frac{2}{3} t_0 + \frac{1}{(1+4\lambda)(3+4\lambda)} \left( 1 - \frac{2\lambda}{\sqrt{y}} \right) + O(f) \]

\[ \leq \frac{2}{3} t_0 + \frac{1}{(1+4\lambda)(3+4\lambda)} \left( 1 - \frac{2\lambda}{\sqrt{y}} \right) + O(f) \]

\[ \frac{2}{3} t_0 + \frac{2e^{-2\lambda L}}{(3+16\lambda+16L^2) - \frac{2\lambda}{\sqrt{y}} (16+32\lambda) + (6+32\lambda+32L^2) L} + O(f) . \]

Now, let $C := (16 + 32\lambda) + (6 + 32\lambda + 32L^2) L, t_0 = \left( \frac{c + \lambda}{(3+16\lambda+16L^2) \lambda} \right)^2$, we get

\[ \text{Var}[X_G] \leq \frac{2e^{-2\lambda L}}{3(3+16\lambda+16L^2)^2} \left( (\varepsilon \lambda L C) \frac{1}{2} + \frac{9 + 48\lambda + 48\lambda^2}{1 - (\varepsilon \lambda L C) \frac{1}{2}} + O(f) \right) \]

\[ \leq \frac{2e^{-2\lambda L}}{3(3+16\lambda+16L^2)^2} \left( (\varepsilon \lambda L C) \frac{1}{2} + \frac{3(\varepsilon \lambda L C) \frac{1}{2}}{1 - (\varepsilon \lambda L C) \frac{1}{2}} + O(f) \right) . \]

Now, recalling that $\varepsilon = \frac{e^{-2\lambda L}}{C} \left( \frac{20(\lambda + \lambda^2)}{9(1+2\lambda)^2} \right)^{\frac{1}{2}}.$
Recall also that under Proposition 2, proved below, under conditions of Theorem 2, we have \( \text{Var}(\bar{w}) = \Theta(1) \) and 
\[
(1 - \delta) \frac{20}{27} \lambda^3 + \frac{20}{27} \lambda^3 (1 + 16\lambda^2) + O(f) \]
and thus,
\[
\frac{\text{Var}(\bar{w})}{\text{Var}(\bar{w})} = \sqrt{\frac{3}{2}} f + O(f^2). \tag{15}
\]

Similarly, we can compute the ratio of mean and variance for \( Y \) (corresponding to unweighted ASTRAL):
\[
\mathbb{E}[Y_G] = \mathbb{E}[\delta_G(ab|cd) - \delta_G(ac|bd)] = 1 - e^{-\lambda} = f + O(f^2) \]
\[
\text{Var}[Y_G] = \text{Var}[\delta_G(ab|cd) - \delta_G(ac|bd)] = \frac{5}{3} f - e^{-2\lambda} = \frac{2}{3} + O(f) \]
and thus,
\[
\frac{\mathbb{E}[Y_G]}{\text{Var}[Y_G]} = \sqrt{\frac{3}{2}} f + O(f^2). \tag{16}
\]

Given Proposition 2, we can use Berry–Esseen theorem to derive
\[
P(\bar{X}_G \leq 0) = P\left( \frac{\sqrt{\lambda}}{\text{Var}(\bar{X}_G)} (\bar{X}_G - \mathbb{E}[\bar{X}_G]) \right) \leq \Phi\left( -\frac{\sqrt{\lambda}}{\text{Var}(\bar{X}_G)} \mathbb{E}[\bar{X}_G] \right) = \Phi\left( -\frac{\sqrt{\lambda}}{\text{Var}(\bar{X}_G)} \mathbb{E}[\bar{X}_G] \right) + O\left( \frac{1}{\sqrt{k}} \right),
\]
where \( \Phi \) denotes CDF of the standard Normal distribution. Since \( k = \Theta(f^{-2}) \),
\[
P(\bar{X}_G \leq 0) = \Phi\left( -\frac{\sqrt{\lambda}}{\text{Var}(\bar{X}_G)} \mathbb{E}[\bar{X}_G] \right) + O(f) \tag{17}
\]
and
\[
P(\bar{Y}_G \leq 0) = \Phi\left( -\frac{\sqrt{\lambda}}{\text{Var}(\bar{Y}_G)} \mathbb{E}[\bar{Y}_G] \right) + O(f). \tag{18}
\]
Combining equations (17) and (18) with (15) and (16), we get
\[ P\left( \sum_{G \in \mathcal{G}} w_G(ab|cd) \leq \sum_{G \in \mathcal{G}} w_G(ac|bd) \right) = \Phi\left( -\sqrt{\frac{3+16\lambda+16\lambda^2}{3+16\lambda+15\lambda^2}} \sqrt{\frac{\sqrt{3}}{2}} f \sqrt{k} \right) + O(f) \]
and
\[ P\left( \sum_{G \in \mathcal{G}} \delta_G(ab|cd) \leq \sum_{G \in \mathcal{G}} \delta_G(ac|bd) \right) = \Phi\left( -\sqrt{\frac{3}{2}} f \sqrt{k} \right) + O(f). \]

As \( f \to 0 \), the interval \( \left( -\sqrt{1 + \frac{4\lambda+4\lambda^2}{3(1+2\lambda)^2}} \sqrt{\frac{\sqrt{3}}{2}} f \sqrt{k}, -\sqrt{\frac{\sqrt{3}}{2}} f \sqrt{k} \right) \) does not shrink because \( \Theta(f \sqrt{k}) = \Theta(1) \). Thus, we have
\[ \Phi\left( -\sqrt{\frac{\sqrt{3}}{2}} f \sqrt{k} \right) - \Phi\left( -\sqrt{1 + \frac{4\lambda+4\lambda^2}{3(1+2\lambda)^2}} \sqrt{\frac{\sqrt{3}}{2}} f \sqrt{k} \right) = \Theta(1) \]
ensuring that
\[ P\left( \sum_{G \in \mathcal{G}} w_G(ab|cd) \leq \sum_{G \in \mathcal{G}} w_G(ac|bd) \right) \leq P\left( \sum_{G \in \mathcal{G}} \delta_G(ab|cd) \leq \sum_{G \in \mathcal{G}} \delta_G(ac|bd) \right). \]

The proof under Proposition 3 is similar. Recall that under Proposition 3, \( \text{Var}[X_{G^*}] = \Theta(f) \) and
\[ \frac{\text{E}[X_{G^*}]}{\sqrt{\text{Var}[X_{G^*}]}} \geq \sqrt{\frac{\sqrt{3}}{2}} \left( 1 - \frac{4\lambda^2}{(1+4\lambda)^2} \right)^{-\frac{1}{2}} f + O(f^2). \] (19)

Given this result, the rest of the proof is similar to the proof under the conditions of Proposition 2, culminating in
\[ P\left( \sum_{G^* \in \mathcal{G}} w_G^*(ab|cd) \leq \sum_{G^* \in \mathcal{G}} w_G^*(ac|bd) \right) \leq \Phi\left( -\left( 1 - \frac{4\lambda^2}{(1+4\lambda)^2} \right)^{-\frac{1}{2}} \sqrt{\frac{\sqrt{3}}{2}} f \sqrt{k} \right) + O(f). \]
\[ \square \]
Placement-based Algorithm

In this section, for a node $v$ in tree $G$, we let $L_v$ denote the set of leaves under $v$.

**Proof of Theorem 3**

**Theorem 3.** Let $S$ be a species tree, $i$ be a species not in $L_S$, $S$ be the set of possible species tree topologies by placing $i$ onto $S$, and $S'$ be the output of Algorithm S1. Then, $W(S', G) = \max_{S \in S} W(S, G)$.

**Proof.** We start with two propositions, proved below.

**Proposition 5.** After each call to ColorLeafSet($L^*, X, T, G, W$) with a $T \neq \emptyset$, $W[T] = \sum_{G \in S} W(T, G)$.

**Proposition 6.** Before calling OptimalTreeDP in line 6 of Algorithm S1, lookup table $W$ contains all tripartitions corresponding to internal nodes of all tree topologies in $S$.

By Proposition 6, all tripartitions corresponding to internal nodes of all tree topologies in $S$ pre-computed. Then, OptimalTreeDP uses a dynamic programming algorithm similar to the one formulated by Mirarab and Warnow 2015 to compute $\arg\max_{S \in S} W(S, G)$. □

**Proposition 5.** After each call to ColorLeafSet($L^*, X, T, G, W$) with a $T \neq \emptyset$, $W[T] = \sum_{G \in S} W(T, G)$.

**Proof.** For a gene tree node $w$ and a color $X$, let $L_w^X$ denote the set of leaves in $L_w$ colored by $X$. For an internal node $w$, let $u, v$ be the children of $w$, $p$ be the parent of $w$ (if $w$ is not the root), and $e$ denote the branch $(w, p)$. For a leaf $i$ and internal node $w$, let $P_{i, w}$ denote path between $i$ and $w$ and $s(P) = 1 - \prod_{e \in P}(1 - s(e))$. For leaves $i, j$, let $m(i, j)$ denote MRCA of $i$ and $j$. Referring back to Table S1, we first establish the connection between recursive formulas of the algorithm and counter definitions.

- When $u_X = \sum_{i \in L^X} e^{-l(P_{i, w})}$, $v_X = \sum_{i \in L^X} e^{-l(P_{i, w})}$,
  
  \[ u_X := \left( u_X + v_X \right) e^{-l(v)} = \sum_{i \in L^X} e^{-l(P_{i, w})} e^{-l(v)} = \sum_{i \in L^X} e^{-l(P_{i, w})}. \]

- When $u_{XX}^X = \sum_{(i, j) \subseteq L^X} e^{-l(P_{i, j})}$, $v_{XX}^X = \sum_{(i, j) \subseteq L^X} e^{-l(P_{i, j})}$,
  
  \[ u_{XX}^X := u_{XX}^X + v_{XX}^X + u_X v_X = \sum_{(i, j) \subseteq L^Y} e^{-l(P_{i, j})} + \sum_{(i, j) \subseteq L^X} e^{-l(P_{i, j})} + \sum_{i \in L^X} \sum_{j \in L^X} e^{-l(P_{i, j})} e^{-l(P_{j, i})} = \sum_{(i, j) \subseteq L^X} e^{-l(P_{i, j})} + \sum_{(i, j) \subseteq L^X} e^{-l(P_{i, j})} + \sum_{i \in L^X} \sum_{j \in L^X} e^{-l(P_{i, j})} e^{-l(P_{j, i})} = \sum_{(i, j) \subseteq L^X} e^{-l(P_{i, j})}. \]

- For $X \neq Y$, when $u_{XY} = \sum_{(i, j) \subseteq L^X \times L^Y} e^{-l(P_{i, j})}$, $v_{XY} = \sum_{(i, j) \subseteq L^X \times L^Y} e^{-l(P_{i, j})}$,
  
  \[ u_{XY}^X := u_{XY}^X + v_{XY}^X + u_X v_Y + u_Y v_X = \sum_{(i, j) \subseteq L^X \times L^Y} e^{-l(P_{i, j})} + \sum_{(i, j) \subseteq L^X \times L^Y} e^{-l(P_{i, j})} + \sum_{(i, j) \subseteq L^X \times L^Y} e^{-l(P_{i, j})} + \sum_{(i, j) \subseteq L^X \times L^Y} e^{-l(P_{i, j})} = \sum_{(i, j) \subseteq L^X \times L^Y} e^{-l(P_{i, j})}. \]
When $u_{XX} = \sum_{(i,j) \in \mathbb{Z}^2} e^{-i(P_{m(i,j,0)})} \prod_{(i',j') \in \mathbb{Z}^2} (1 - s(\hat{e}^{i,j}))$, $v_{XX} = \sum_{(i,j) \in \mathbb{Z}^2} e^{-i(P_{m(i,j,0)})} \prod_{(i',j') \in \mathbb{Z}^2} (1 - s(\hat{e}^{i,j}))$, and $X \neq Y$, similarly,

$$w_{XY} := (u_{XX} + v_{XY} + u_{XY} v_{XY}) (1 - s(\hat{e}^{i,j})) = \sum_{(i,j) \in \mathbb{Z}^2} e^{-i(P_{m(i,j,0)})} (1 - s(\hat{e}^{i,j})) .$$
Similarly,

\[ u_{X|Y|Z} = \sum_{h \in L_X} \sum_{k \in L_Y} w_G(hk|ij), u_{X|Z|Y} = \sum_{i \in L_X} \sum_{j \in L_Y} w_G(hi|jk), u_{X|X|Y} = \sum_{j \in L_X} \sum_{k \in L_Y} w_G(hi|jk), \]

Then, the value returned by \( Q \) at the end of procedure \( \text{UpdateCounters} \), \( \sum_{G \in \mathcal{G}} w_G(X|Y|Z) = \sum_{(h,i) \in L_G^X} \sum_{(j,k) \in L_G^Y} w_G(hi|jk). \]

Also,

\[ u_{Y|Z|X} = \sum_{(h,i) \in L_G^X} \sum_{(j,k) \in L_G^Y} e^{-(P_{h,i})-(P_{j,k})} \]

\[ \sum_{(h,i) \in L_G^X} \sum_{(j,k) \in L_G^Y} w_G(hi|jk). \]

Notice that above cases count exactly once all quartets \( hi|jk \) for all leaf nodes \( h,i \) colored \( X \), \( j \) colored \( Y \), \( k \) colored \( Z \) such that MRCA of \( h,i,j,k \) is \( w \); namely,

\[ w_{X|Y|Z} = \sum_{(h,i) \in L_G^X} \sum_{(j,k) \in L_G^Y} w_G(hi|jk). \]

We define \( I(G) \) to be the set of internal nodes of gene tree \( G \) and \( L_G^X \) be the set of leaves of gene tree \( G \) with color \( X \). It is trivial to verify that at the end of procedure \( \text{UpdateCounters} \), \( \sum_{G \in I(G)} w_{X|Y|Z} = \sum_{(h,i) \in L_G^X} \sum_{(j,k) \in L_G^Y} w_G(hi|jk). \]

Thus, \( Q \) returned by \( \text{UpdateCounters} \) satisfies:

\[ Q = \sum_{G \in \mathcal{G}} \left( \sum_{(h,i) \in L_G^X} w_G(hi|jk) + \sum_{(j,k) \in L_G^Y} w_G(hi|jk) + \sum_{(h,i) \in L_G^X} w_G(hi|jk) \right). \]

For tripartition \( T = A|B|C \), note that by assumption, before the call, all the gene tree leaves are colored such that recoloring \( C \) by \( X \) would produce a coloring that matches \( T \). Thus, at the end of the call to \( \text{ColorLeafSet} \), for each gene tree \( G \), we have \( A \cap L_G = L_G^A, B \cap L_G = L_G^B, \) and \( C \cap L_G = L_G^C \).

Then, the value returned by \( \text{UpdateCounters} \) satisfies:

\[ Q = \sum_{G \in \mathcal{G}} W(A|B|C,G). \]
It can be easily verified that after each call to \( \text{ColorLeafSet}(L^*,X,T,G,W) \), the species tree tripartition \( T \) matches the coloring of all gene trees as required by conditions of (20), concluding \( W[T]=Q=\sum_{c\in\mathcal{G}} W(T,G) \).

**Proposition 6.** Before calling \( \text{OptimalTreeDP} \) in line 6 of Algorithm S1, lookup table \( W \) contains all tripartitions corresponding to internal nodes of all tree topologies in \( S \).

**Proof.** Each \( S \in \mathcal{S} \) places \( i \) above a different node \( w \) of \( S \) creating a new node corresponding to tripartition \( L_w[\{i\}]\mathcal{L}_S - L_w \) covered in line 24. Besides new nodes, each existing internal node \( w \) of \( S \) will correspond to a different tripartition after placing \( i \) onto \( S \) depending on the relative location of \( w \) and \( i \). Let \( u,v \) denote the larger and the smaller child of \( w \). Node \( w \) corresponds to \( L_u[\{i\}]\mathcal{L}_v|\mathcal{L}_S - L_w \) if \( i \) is under \( u \), corresponds to \( \{i\} \cup L_u|\mathcal{L}_S - L_w \) if \( i \) is under \( v \), and corresponds to \( L_u|\mathcal{L}_v[\{i\}] \cup \mathcal{L}_S - L_w \) if \( i \) is above \( w \). All three cases for each node \( w \) is covered in lines 20–22. □

**Proof of Theorem 4**

**Theorem 4.** If there exists a species tree topology \( S^* \) satisfying that for each quartet subtree \( ab|cd \),

\[
\sum_{c\in\mathcal{G}} w(ab|cd) > \max \left( \sum_{c\in\mathcal{G}} w(ac|bd), \sum_{c\in\mathcal{G}} w(ad|bc) \right),
\]

then the output of Algorithm S2 will be \( S^* \).

**Proof.** We start with a Corollary 1 of Theorem 3

**Corollary 1.** Assuming (6), if \( S \) is compatible with the true tree \( S^* \), then \( S' \) is compatible with \( S^* \).

By induction, \( W_i \) in line 8 of Algorithm S2 should contain all tripartitions of \( S^* \), as at that time \( S_i = S^* \) by Corollary 1. Consequentially, the output of Algorithm S2 must also be \( S^* \). □

**Proof of Proposition 4**

**Proposition 4.** The time complexity of Algorithm S2 is \( O(kHn^2 \log n) \).

**Proof.** We begin with a proposition and a corollary.

**Proposition 7.** Procedure \( \text{ColorNode} \) on any species tree node \( w \) takes \( O(kH|L_w|\log |L_w|) \) time.

**Proof (sketch) of Proposition 7.** We can prove this proposition by induction. For an internal node \( w \) with larger child \( u \) and smaller child \( v \), if for some constant \( C \geq \frac{1}{\log^2} \), \( \text{ColorNode} \) on \( u \) calls \( \text{UpdateCounters} \) at most \( Ck|L_u|\log |L_u| + 1 \) times and \( \text{ColorNode} \) on \( v \) calls \( \text{UpdateCounters} \) at most \( Ck|L_v|\log |L_v| + 1 \) times, then \( \text{ColorNode} \) on \( w \) calls \( \text{UpdateCounters} \) at most

\[
Ck|L_u|\log |L_u| + 1 + Ck|L_u|\log |L_u| + 1 + 3k(|L_u| + 1) \\
\leq Ck|L_u|\log |L_u| + 1 + Ck|L_u|\log \frac{|L_u|}{2} + 1 + 6k|L_u| \\
\leq Ck|L_u|\log |L_u| + 1 + Ck|L_u|\log |L_u| + 1 - Ck|L_u|\log 2 + 6k|L_u| \\
\leq Ck|L_u|\log |L_u| + 1 + (6 - C\log 2)k|L_u| \\
\leq Ck|L_u|\log |L_u| + 1 \text{ times.}
\]
It is easy to verify that each UpdateCounters takes $O(H_G)$ time where $H_G$ is the height of the gene tree, and thus ColorNode on node $w$ takes $O(kH|L_w|\log |L_w|)$ time.

**Corollary 2** (Corollary of Proposition 7). For any tree topology $S$ with $n$ species, the Place procedure on $S$ takes $O(kH\log n)$ time.

**NaivePlacement** of taxon set $T$ makes $r(|T|−3)$ calls to Place, each of which takes $O(kH|T|\log |T|)$ time. Thus, NaivePlacement takes $O(kH|T|^2\log |T|)$ time and when $T=L_S$ and $r=O(1)$, $O(rkH|T|^2\log |T|)=O(n^2kH\log n)$. □

---

**Proofs of Theorems 6 and Theorem 5**

**Theorem 6.** Under the conditions of Theorem 4, the DAC Algorithm S3 will output $S^*$.

**Proof.** By Theorem 4, $S_i$ in line 5 of Algorithm S3 are compatible with $S^*$. With Corollary 1, by induction, each $S_i$ in line 21 of Algorithm S3 is compatible with $S^*$. Consequentially, $W_i$ in line 26 contain all tripartitions of $S^*$, as at that time $S_i^*=S^*$, and the output of Algorithm S3 must also be $S^*$. □

**Theorem 5.** When the inequality condition in Theorem 4 is satisfied, then the time complexity of the DAC algorithm is $O(n^{1.5+t}kH)$ with arbitrarily high probability.

**Proof (sketch).** From the inequality (6), we can trivially deduce that $S^*$ is the species tree topology that maximizes the weighted quartet score, and each $S_i$ in line 5 of Algorithm S3 is compatible to $S^*$. Also, each $C_e$ in line 15 of Algorithm S3 equals the set of species under the edges coming off of the internal nodes on the path of $S^*$ corresponding to $e$.

We now introduce a proposition

**Proposition 8.** With high probability, $\max_{e \in E_{S_i^*}} |C_e| \leq 2\sqrt{n}\log n + O(\sqrt{n})$.

**Proof.** For each pair of nodes $u, v$ of $S^*$, let $C_{u,v}:= \{x : x \in L_S, u \text{ is not on } P_{x,v} \text{ and } v \text{ is not on } P_{x,u}\}$. It is easy to verify that for every $e$ of $S_i$, $C_e=C_{u,v}$ for some nodes $u,v$ of $S^*$. For every $u$ and $v$ that are sufficiently apart so that $C_{u,v}$ has $2\sqrt{n}\log n + \omega(\sqrt{n})$ elements and a random $T_i$ in line 4 of Algorithm S3,

$$P(C_{u,v} \cap T_i = \emptyset) \leq \left(1 - \frac{1}{\sqrt{n}}\right)^{2|C_{u,v}|} \leq e^{-\frac{2|C_{u,v}|}{n}} = o\left(\frac{1}{n^2}\right).$$

By union bound, the probability that there exists a pair of nodes $u, v$ of $S^*$ such that $|C_{u,v}| \geq 2\sqrt{n}\log n + \omega(\sqrt{n})$ and $C_{u,v} \cap T_i = \emptyset$ is $o(1)$. Since, by definition, $C_e \cap T_i = \emptyset$ for every $C_e$, with high probability, there exists no $C_e$ having $2\sqrt{n}\log n + \omega(\sqrt{n})$ elements.

Since $|T_i| \sim \text{Binomial}(n, \frac{1}{\sqrt{n}})$, with high probability $|T_i| = O(\sqrt{n})$ and calling NaivePlacement on line 5 takes $O(n^{1.5}kH\log n)$ time. It is easy to confirm that $C_\emptyset = \emptyset$ and every call to Place takes as input a species tree topology of $O(\sqrt{n}\log n)$ species with high probability. Thus, with high probability, each call to Place takes $O(\sqrt{n}kH\log^2 n\log n)$ time and all $O(n)$ calls to Place takes $O(n^{1.5}kH\log^2 n\log n)$ time. Therefore, the time complexity of the DAC algorithm is $O(n^{1.5}kH\log^2 n\log n)=O(n^{1.5+t}kH)$ with high probability. □