Materials and Methods

1 Notation

1.1 Phylogenies

A phylogeny is a rooted leaf-labeled tree $\mathcal{T} = (V, E, [n], r; \delta)$ where V is the set of vertices, E is the set of edges, $L = [n] = \{0, \ldots, n-1\}$ is the set of leaves, r is the root, and $\delta : E \to (0, +\infty)$ is the branch length function. It is further assume that all internal nodes in $\mathcal{T}$ have degree 3 except for the root r which has degree 2. For two leaves $a, b \in [n]$, the set of edges on the unique path between a and b is denoted by $\mathrm{Path}(a, b)$. The tree metric corresponding to the phylogeny $\mathcal{T} = (V, E, [n], r; \delta)$ is denoted by $(\delta(a, b))_{a,b\in[n]}$, that is,

$$\delta(a, b) = \sum_{e\in \mathrm{Path}(a,b)} \delta(e),$$

for all leaves $a, b \in [n]$. The quantity $\delta(u, v)$ is extended to all vertices $u, v \in V$ in the obvious way. Let $\mathbf{Y}_n$ be the set of all such phylogenies on n leaves and denote $\mathbf{Y} = \{\mathbf{Y}_n\}_{n\geq 1}$. Finally, the set of all phylogenies $\mathcal{T} = (V, E, [n], r; \delta)$ where further δ is ultrametric, that is, for all $v \in V$ it holds that $\delta(v, x) = \delta(v, y) \equiv \delta_v$, for all leaves x, y below v is denoted by $\mathbf{UY}$. This is the molecular clock case.

1.2 Model of molecular sequence evolution

A standard model of evolution for molecular sequences on a phylogeny $\mathcal{T} = (V, E, [n], r; \delta)$ is a Markov model on a tree (MMT). Let $\Phi = \{\mathtt{A}, \mathtt{G}, \mathtt{C}, \mathtt{T}\}$. For each edge $e \in E$, a 4×4 stochastic matrix $M^e = (M^e_{ij})_{i,j\in\Phi}$ is given, with fixed stationary distribution $\pi = (\pi_i)_{i\in\Phi} > 0$. An MMT associates a state s_v to each vertex v in V as follows: pick a state for the root r according to π; moving away from the root, choose a state for each vertex v independently according to the distribution $(M^e_{s_u,j})_{j\in\Phi}$, with $e = (u, v)$ where u is the parent of v. A common MMT used in

phylogenetics is the General Time-Reversible (GTR) model. Let Q be a 4×4 rate matrix, that is, $Q_{ij} > 0$ for all $i \neq j$ and $\sum_{j \in \Phi} Q_{ij} = 0$, for all $i \in \Phi$. Assume Q is reversible with respect to π, that is, $\pi_i Q_{ij} = \pi_j Q_{ji}$, for all $i, j \in \Phi$. The GTR model with rate matrix Q is the MMT with transition matrices $M^e = e^{\delta(e)Q}$, for all $e \in E$. By the reversibility assumption, Q has 4 real eigenvalues $0 = \Lambda_1 > \Lambda_2 \geq \cdots \geq \Lambda_4$. The matrix Q is normalized by fixing $\Lambda_2 = -1$. The vector of states on the vertices $W \subseteq V$ is denote by s_W. In particular, $s_{[n]}$ are the states at the leaves. GTR models include as special cases many standard models such as the Jukes-Cantor (JC) where $\pi = (1/4, \ldots, 1/4)$ and $Q_{ij} = \frac{1}{4}$ if $i \neq j$. The following more general model will also be used.

Example 1 (Tamura-Nei Model)**.** In the Tamura-Nei model (*S1*), the state space is

$$\Phi = \{\mathtt{A}, \mathtt{G}, \mathtt{C}, \mathtt{T}\},$$

with stationary distribution $\pi = (\pi_{\mathtt{A}}, \pi_{\mathtt{G}}, \pi_{\mathtt{C}}, \pi_{\mathtt{T}})$. The rate matrix is given by

$$Q = \begin{pmatrix} - & \alpha_R \pi_{\mathtt{G}}/\pi_R + \beta\pi_{\mathtt{G}} & \beta\pi_{\mathtt{C}} & \beta\pi_{\mathtt{T}} \\ \alpha_R \pi_{\mathtt{A}}/\pi_R + \beta\pi_{\mathtt{A}} & - & \beta\pi_{\mathtt{C}} & \beta\pi_{\mathtt{T}} \\ \beta\pi_{\mathtt{A}} & \beta\pi_{\mathtt{G}} & - & \alpha_Y \pi_{\mathtt{T}}/\pi_Y + \beta\pi_{\mathtt{T}} \\ \beta\pi_{\mathtt{A}} & \beta\pi_{\mathtt{G}} & \alpha_Y \pi_{\mathtt{C}}/\pi_Y + \beta\pi_{\mathtt{C}} & - \end{pmatrix}$$

where $\pi_R = \pi_{\mathtt{A}} + \pi_{\mathtt{G}}$, $\pi_T = \pi_{\mathtt{C}} + \pi_{\mathtt{T}}$, $\beta, \alpha_R, \alpha_Y \geq 0$, and the diagonal is obtained by the condition that the rows sum to 0. The rates of transitions and transversions are respectively (*S2*)

$$T_s = 2\alpha_R \pi_{\mathtt{A}} \pi_{\mathtt{G}}/\pi_R + 2\alpha_Y \pi_{\mathtt{C}} \pi_{\mathtt{T}}/\pi_Y + \beta(2\pi_{\mathtt{A}}\pi_{\mathtt{G}} + 2\pi_{\mathtt{C}}\pi_{\mathtt{T}}),$$

and

$$T_v = 2\beta\pi_R\pi_Y.$$

The transition to transversion ratio is denoted by $R = T_s/T_v$. By checking by $(\pi_{\mathtt{G}}, -\pi_{\mathtt{A}}, 0, 0)$ and $(0, 0, \pi_{\mathtt{T}}, -\pi_{\mathtt{C}})$ are right eigenvectors and using the fact that the trace is the sum of eigenvalues, it can be shown that the eigenvalues of Q are 0, $-\beta$, $-\beta - \alpha_R$, $-\beta - \alpha_Y$. Therefore, to normalize

Q as above, one must divide the rate matrix by β. The Tamura-Nei model includes as special cases (see e.g. (*S2*)): the Jukes-Cantor model ($\pi = (1/4, 1/4, 1/4, 1/4)$, $\alpha_R = \alpha_Y = 0$); the Kimura two-parameter model ($\pi = (1/4, 1/4, 1/4, 1/4)$, $\alpha_R = \alpha_Y$); the F84 model ($\alpha_R = \alpha_Y$); the HKY model ($\alpha_R/\alpha_Y = \pi_R/\pi_Y$).

Remark 1 (Biological Convention)**.** The normalization of Q above differs from standard biological convention where it is assumed that the total rate of change per unit time at stationarity is 1, that is,

$$\sum_i \pi_i Q_{ii} = -1.$$

See e.g. (*S2*). Let $-\lambda_Q$ denote the largest negative eigenvalue under this convention. Then, the critical branch length (see main text) is given by the solution to

$$2e^{-2\lambda_Q g_Q^*} = 1.$$

For instance, in the Tamura-Nei model, it can be shown that under the convention above (that is, $T_s + T_v = 1$) it must be that

$$\lambda_Q = \beta = \frac{1}{2\pi_R \pi_Y (1 + R)},$$

and hence

$$g_Q^* = \pi_R \pi_Y (1 + R) \ln 2.$$

In the special cases of the Jukes-Cantor and Kimura two-parameter models one has

$$g_{\mathrm{JC}}^* = \frac{3}{8} \ln 2,$$

and

$$g_{\mathrm{K2P}}^* = \frac{R + 1}{4} \ln 2.$$

1.3 Phylogenetic reconstruction

A standard assumption in molecular evolution is that each site in a sequence evolves independently according to a GTR model. Because of the reversibility assumption, the root of the phylogeny cannot be identified and phylogenies are reconstructed up to their root. Let $\widetilde{\mathbf{Y}} = \{\widetilde{\mathbf{Y}}_n\}_{n\geq 1}$ be a subset of phylogenies and $\widetilde{\mathbf{Q}}$ be a subset of rate matrices on 4 states. Let $\mathcal{T} = (V, E, [n], r; \delta) \in \widetilde{\mathbf{Y}}$. If $T = (V, E, [n], r)$ is the rooted tree underlying $\mathcal{T}$, denote by $T_-[\mathcal{T}]$ the tree T where the root is removed: that is, the two edges adjacent to the root are replaced by a single edge. Denote by $\mathbf{T}_n$ the set of all leaf-labeled trees on n leaves with internal degrees 3 and let $\mathbf{T} = \{\mathbf{T}_n\}_{n\geq 1}$. A phylogenetic reconstruction algorithm is a collection of maps $\mathcal{A} = \{\mathcal{A}_{n,k}\}_{n,k\geq 1}$ from sequences $(s^i_{[n]})_{i=1}^k \in (\Phi^{[n]})^k$ to leaf-labeled trees $T \in \mathbf{T}_n$. Only algorithms computable in time polynomial in n and k are considered. Let $k(n)$ be an increasing function of n. Say that $\mathcal{A}$ solves the phylogenetic reconstruction problem on $\widetilde{\mathbf{Y}} \otimes \widetilde{\mathbf{Q}}$ with sequence length $k = k(n)$ if for all $\varepsilon > 0$, there is $n_0 \geq 1$ such that for all $n \geq n_0$, $\mathcal{T} \in \widetilde{\mathbf{Y}}_n$, $Q \in \widetilde{\mathbf{Q}}$,

$$\mathbf{P}\left[\mathcal{A}_{n,k(n)}\left((s^i_{[n]})_{i=1}^{k(n)}\right) = T_-[\mathcal{T}]\right] \geq 1 - \varepsilon,$$

where $(s^i_{[n]})_{i=1}^{k(n)}$ are i.i.d. samples from the GTR model on $\mathcal{T}$ with rate matrix Q.

1.4 Distance methods

Let $(s^i_a)_{i=1}^k, (s^i_b)_{i=1}^k \in \Phi^k$ be the sequences at $a, b \in [n]$. For $\upsilon_1, \upsilon_2 \in \Phi$, define the correlation matrix between a and b by

$$\widehat{F}^{ab}_{\upsilon_1,\upsilon_2} = \frac{1}{k}\sum_{i=1}^{k} \mathbf{1}\{s^i_a = \upsilon_1, s^i_b = \upsilon_2\},$$

that is, the proportion of sites at which a is υ_1 and b is υ_2. Let $\widehat{F}^{ab} = (\widehat{F}^{ab}_{\upsilon_1,\upsilon_2})_{\upsilon_1,\upsilon_2\in\Phi}$ be the corresponding matrix. A phylogenetic reconstruction algorithm is said distance-based if it depends on the data $(s^i_{[n]})_{i=1}^k \in (\Phi^{[n]})^k$ only through the correlation matrices $\{\widehat{F}^{ab}\}_{a,b\in[n]}$.

The previous definition takes a general view of distance-based methods: any method that uses only pairwise sequence comparisons. In practice, most distance-based approaches actually use a specific distance estimator, that is, a function of $\widehat{F}^{ab}$ that converges to $\delta(a,b)$ in probability as $n \to +\infty$.

2 Molecular clock case

In the rest of this section, the rate matrix Q is the Jukes-Cantor matrix and a molecular clock is assumed to hold.

2.1 Preliminaries

Under a molecular clock, it is known that it suffices to consider uncorrected distances (*S3*). An uncorrected distance estimate between leaves a and b is obtained by letting

$$\hat{\delta}_{\mathrm{u}}(a,b) = \text{proportion of transversions between } a \text{ and } b. \tag{1}$$

Note that one can write $\hat{\delta}_{\mathrm{u}}(a,b) = (1 - \widehat{\Theta}(a,b))/2$, where

$$\widehat{\Theta}(a,b) = \nu^\top \widehat{F}^{ab} \nu = \frac{1}{k} \sum_{i=1}^{k} [\mathbf{1}\{s_a^i \leftrightarrow s_b^i \text{ is a transition}\} - \mathbf{1}\{s_a^i \leftrightarrow s_b^i \text{ is a transversion}\}],$$

with $\nu = (1, 1, -1, -1)$, a right eigenvector of the rate matrix Q corresponding to the second eigenvalue $\Lambda_2 = -1$. Eq. (1) is indeed a legitimate distance estimator. Note that $\mathbf{E}[\widehat{F}_{ij}^{ab}] = \pi_i \left(e^{-\delta(a,b)Q}\right)_{ij}$. Hence,

$$\begin{aligned}
\Theta(a,b) &= \mathbf{E}[\widehat{\Theta}(a,b)] \\
&= \mathbf{E}\left[\nu^\top \widehat{F}^{ab} \nu\right] \\
&= \sum_{i \in \Phi} \nu_i \sum_{j \in \Phi} \pi_i \left(e^{-\delta(a,b)Q}\right)_{ij} \nu_j \\
&= \sum_{i \in \Phi} \nu_i (\pi_i e^{-\delta(a,b)} \nu_i) \\
&= e^{-\delta(a,b)}.
\end{aligned}$$

Therefore,

$$\delta_{\mathrm{u}}(a,b) = \mathbf{E}[\hat{\delta}_{\mathrm{u}}(a,b)] = \frac{1-e^{-\delta(a,b)}}{2} = \text{expected proportion of transversions between } a \text{ and } b.$$

For the rest of this section, it will be convenient to work primarily with $\widehat{\Theta}$ rather than $\hat{\delta}_{\mathrm{u}}$. Note that $\widehat{\Theta}$ can be written in a different, but equivalent, form. For $a \in [n]$ and $i = 1, \ldots, k$, let $\sigma_a^i = \nu_{s_a^i}$. In words, purines are denoted by $+1$ and pyrimidines are denoted by -1. Then (1) is equivalent to

$$\widehat{\Theta}(a,b) = \frac{1}{k}\sum_{i=1}^{k} \sigma_a^i \sigma_b^i. \tag{2}$$

For future reference, the variances and covariances of $\widehat{\Theta}$ are also computed. Note that

$$\mathrm{Var}[\widehat{\Theta}(a,b)] = \frac{1}{k}\mathrm{Var}[\sigma_a^1\sigma_b^1] = \frac{1}{k}(\mathbf{E}[(\sigma_a^1\sigma_b^1)^2] - \mathbf{E}[\sigma_a^1\sigma_b^1]^2) = \frac{1}{k}(1-e^{-2\delta(a,b)}).$$

Similarly,

$$\begin{aligned}
\mathrm{Cov}[\widehat{\Theta}(u,v),\widehat{\Theta}(x,y)] &= \mathbf{E}[\widehat{\Theta}(u,v)\widehat{\Theta}(x,y)] - \mathbf{E}[\widehat{\Theta}(u,v)]\mathbf{E}[\widehat{\Theta}(x,y)] \\
&= \frac{1}{k^2}\left(k\mathbf{E}[\sigma_u^1\sigma_v^1\sigma_x^1\sigma_y^1] + k(k-1)\mathbf{E}[\sigma_u^1\sigma_v^1]\mathbf{E}[\sigma_x^1\sigma_y^1]\right) - \mathbf{E}[\sigma_u^1\sigma_v^1]\mathbf{E}[\sigma_x^1\sigma_y^1] \\
&= \frac{1}{k}\left(\mathbf{E}[\sigma_u^1\sigma_v^1\sigma_x^1\sigma_y^1] - \mathbf{E}[\sigma_u^1\sigma_v^1]\mathbf{E}[\sigma_x^1\sigma_y^1]\right),
\end{aligned}$$

where the last expression depends on how the tree splits the set of leaves $\{u,v,x,y\}$. Note that $\mathbf{E}[\sigma_u^1\sigma_v^1\sigma_x^1\sigma_y^1]$ is equal to $\mathbf{E}[\sigma_u^1\sigma_v^1]\mathbf{E}[\sigma_x^1\sigma_y^1]$ if $uv|xy$ and similarly for the other splits. Hence one gets

$$\mathrm{Cov}[\widehat{\Theta}(u,v),\widehat{\Theta}(x,y)] = \begin{cases} 0, & \text{if } uv|xy, \\ \frac{1}{k}(e^{-\delta(u,x)-\delta(v,y)} - e^{-\delta(u,v)-\delta(x,y)}), & \text{if } ux|vy, \\ \frac{1}{k}(e^{-\delta(u,y)-\delta(v,x)} - e^{-\delta(u,v)-\delta(x,y)}), & \text{if } uy|vx. \end{cases}$$

Finally the variance-covariance matrix of the distance matrix in this case is obtained by noticing

$$\mathrm{Var}[\hat{\delta}_{\mathrm{u}}(a,b)] = \frac{1}{4}\mathrm{Var}[\widehat{\Theta}(a,b)] \qquad \mathrm{Cov}[\hat{\delta}_{\mathrm{u}}(u,v),\hat{\delta}_{\mathrm{u}}(x,y)] = \frac{1}{4}\mathrm{Cov}[\widehat{\Theta}(u,v),\widehat{\Theta}(x,y)]. \tag{3}$$

Let $e = (x,y) \in E$ and assume that x is closest to r (in number of edges). Define

$$R_r(e) = \left(1-\theta_e^2\right)\Theta(r,y)^{-2},$$

where $\Theta(r, y) = e^{-\delta(r,y)}$ and $\theta_e = e^{-\delta(e)}$. The following ancestral state estimator was studied by Mossel and Peres (*S4*). Let w be a convex combination over the leaves, that is, $w(x) \geq 0$ for all $x \in [n]$ and $\sum_{x\in[n]} w(x) = 1$. Consider the root-state estimator

$$S = \sum_{x\in[n]} \frac{w(x)\sigma_x}{\Theta(r, x)}.$$

Then, one has $\mathbf{E}[S] = 0$. Moreover, S is a conditionally unbiased estimator of the state σ_r at the root, that is, the expectation of S given that the root has state σ_r is itself σ_r. There exists an elegant formula for the variance of S, namely,

$$\mathrm{Var}[S] = 1 + K_{r,w},$$

where

$$K_{r,w} = \sum_{e\in E} R_r(e)w(e)^2,$$

and $w(e)$ is the sum of all $w(x)$s over leaves below edge e.

2.2 Two Predictions

In this section, the best distance estimator between two clades is computed under two stochastic models of error for the distances. In the full Markov model, the distribution of the distance estimates is derived from the distribution described in the previous section. In particular, the full variance-covariance matrix of the distance matrix is used. On the other hand, in the independent-error model, it is assumed that the distance estimates are independent with variance as computed in (3).

2.2.1 Independent-error model

Consider first the independent-error model. Let A, B be subsets of leaves corresponding to two disjoint subtrees of T with respective most recent common ancestor (MRCA) a^*, b^*. One seeks

the choice of leaf weights $w(a)$, $a \in A$, and $w(b)$, $b \in B$, that minimizes the variance of the estimator

$$\bar{\delta}_{\mathrm{u}}(A,B) = \sum_{a\in A}\sum_{b\in B} w(a)w(b)\hat{\delta}_{\mathrm{u}}(a,b),$$

where $\sum_{a\in A} w(a) = 1$, and $w(a) \geq 0$, $\forall a \in A$, and similarly for B. This implies

$$\mathbf{E}[\bar{\delta}_{\mathrm{u}}(A,B)] = \mathbf{E}\left[\sum_{a\in A}\sum_{b\in B} w(a)w(b)\hat{\delta}_{\mathrm{u}}(a,b)\right] = \delta_{\mathrm{u}}(A,B),$$

where, using the molecular clock assumption, $\delta_{\mathrm{u}}(A,B) = \delta_{\mathrm{u}}(a,b)$, $\forall a \in A, b \in B$. Similarly let $\delta(A,B) = \delta(a,b)$, $\forall a \in A, b \in B$.

Using independence, (3), and the molecular clock assumption, one has

$$\begin{aligned}
\mathrm{Var}[\bar{\delta}_{\mathrm{u}}(A,B)] &= \mathrm{Var}\left[\sum_{a\in A}\sum_{b\in B} w(a)w(b)\hat{\delta}_{\mathrm{u}}(a,b)\right] \\
&= \sum_{a\in A}\sum_{b\in B} w(a)^2 w(b)^2 \mathrm{Var}\left[\hat{\delta}_{\mathrm{u}}(a,b)\right] \\
&= \frac{1}{4k}(1-e^{-2\delta(A,B)})\sum_{a\in A} w(a)^2 \sum_{b\in B} w(b)^2.
\end{aligned}$$

By standard optimization techniques, the minimum of the above expression is attained for $w(a) = 1/|A|$, $\forall a \in A$, and $w(b) = 1/|B|$, $\forall b \in B$.

2.2.2 Full Markov model

Consider now the full Markov model. Let $T_A = (V_A, E_A)$ be the subtree corresponding to A and similarly for B. For $a \in A$, let $|a|_A = |a|_{a^*}$ and $\Theta_A = \Theta(a^*, a)$, where $|a|_{a^*}$ is the number of branches between a and the most recent common ancestor (MRCA) a^* of A, and similarly for B. Again, consider the inter-clade distance estimator

$$\bar{\delta}_{\mathrm{u}}(A,B) = \sum_{a\in A}\sum_{b\in B} w(a)w(b)\hat{\delta}_{\mathrm{u}}(a,b).$$

In particular, one has $\mathbf{E}[\bar{\delta}_{\mathrm{u}}(A,B)] = \delta_{\mathrm{u}}(a,b)$ as in the independent-error model. It will be convenient to work with $\overline{\Theta}_{\mathrm{u}}(A,B)$ rather than $\bar{\delta}_{\mathrm{u}}(A,B)$. Let

$$\overline{\Theta}_{\mathrm{u}}(A,B) \;=\; \sum_{a\in A}\sum_{b\in B} w(a)w(b)\widehat{\Theta}(a,b).$$

Note that $\mathrm{Var}[\bar{\delta}_{\mathrm{u}}(A,B)] = \frac{1}{4}\mathrm{Var}[\overline{\Theta}_{\mathrm{u}}(A,B)]$ so it suffices to compute the latter. Note the following observation:

$$\begin{aligned}
\overline{\Theta}_{\mathrm{u}}(A,B) &= \Theta_A\Theta_B \sum_{a\in A}\sum_{b\in B} w(a)w(b)\Theta_A^{-1}\Theta_B^{-1}\widehat{\Theta}(a,b)\\
&= \Theta_A\Theta_B \frac{1}{k}\sum_{i=1}^{k}\left(\sum_{a\in A}\frac{w(a)\sigma_a^i}{\Theta_A}\right)\left(\sum_{b\in B}\frac{w(b)\sigma_b^i}{\Theta_B}\right)\\
&= \Theta_A\Theta_B \frac{1}{k}\sum_{i=1}^{k} S_A^i S_B^i,
\end{aligned}$$

where

$$S_A^i = \sum_{a\in A}\frac{w(a)\sigma_a^i}{\Theta_A},$$

is a linear ancestral estimator at a^*, and similarly for B.

Using the expression for the variance of S_A, S_B given in the previous section one gets

$$\begin{aligned}
\mathrm{Var}[\overline{\Theta}_{\mathrm{u}}(A,B)] &= \frac{\Theta_A^2\Theta_B^2}{k}\mathrm{Var}[S_A^1 S_B^1]\\
&= \frac{\Theta_A^2\Theta_B^2}{k}\left[\mathbf{E}[(S_A^1)^2(S_B^1)^2] - \mathbf{E}[S_A^1 S_B^1]^2\right]\\
&= \frac{\Theta_A^2\Theta_B^2}{k}\left[\mathbf{E}[\mathbf{E}[(S_A^1)^2(S_B^1)^2 \mid \sigma_{a^*},\sigma_{b^*}]] - e^{-2\delta(a^*,b^*)}\right]\\
&= \frac{\Theta_A^2\Theta_B^2}{k}[(1+K_{a^*,w})(1+K_{b^*,w}) - e^{-2\delta(a^*,b^*)}],
\end{aligned}$$

(where $\mathbf{E}[(S_A^1)^2(S_B^1)^2 \mid \sigma_{a^*},\sigma_{b^*}]$ is the conditional expectation of $(S_A^1)^2(S_B^1)^2$ given that the states at a^* and b^* are σ_{a^*} and σ_{b^*}). This expression depends on w only through $K_{a^*,w}$ and $K_{b^*,w}$. Hence to minimize the variance of $\bar{\delta}_{\mathrm{u}}(A,B)$ it suffices to solve the following convex

quadratic programs

$$\begin{array}{l} \min_w \sum_{e\in E_A} R_{a^*}(e) w(e)^2 \\ \text{s.t.} \left\{ \begin{array}{l} \sum_{a\in A} w(a) = 1, \\ w(a) \geq 0,\ \forall a \in A, \end{array} \right. \end{array}$$

and similarly for B.

A particularly good solution is obtained by setting $w(a) = 2^{-|a|_A}$, $\forall a \in A$ and $w(b) = 2^{-|b|_B}$, $\forall b \in B$. The power of this choice of weights becomes more apparent when one considers a complete binary tree where for all edges e it holds that $\delta(e) = g < g^{**} = \ln\sqrt{2}$. (The value of g^{**} is different than the value of g^*_{JC} in the main text because of the normalization implied by fixing $\Lambda_2 = -1$. Up to this re-parametrization, the two values are equivalent.) Indeed, letting $h_A = \max_{a\in A} |a|_A$ and summing over the levels starting at the root,

$$\begin{aligned} K_{a^*,w} &= \sum_{i=0}^{h_A-1} 2^{h_A-i} \left\{ (1-e^{-2g}) e^{2(h_A-i)g} (2^{-(h_A-i)})^2 \right\} \\ &= (1-e^{-2g}) \sum_{j=1}^{h_A} e^{2jg} e^{-j(\ln 2)} \\ &= (1-e^{-2g}) \sum_{j=1}^{h_A} e^{-2j(g^{**}-g)} \\ &= (1-e^{-2g}) e^{-2(g^{**}-g)} \frac{1-e^{-2h_A(g^{**}-g)}}{1-e^{-2(g^{**}-g)}} \\ &\equiv \Upsilon(g, h_A), \end{aligned} \tag{4}$$

which is bounded uniformly in h_A (for fixed g), and similarly for B. (A similar bound holds for general clock-like topologies as long as for all edges $\delta(e) \leq g$.) That is, the accuracy of the estimator does not deteriorate as one reaches deeper and deeper into the phylogeny. To see the significance of (4), note that it gives a signal-to-noise ratio for $\overline{\Theta}_{\mathrm{u}}(A, B)$ of

$$\frac{\mathbf{E}[\overline{\Theta}_{\mathrm{u}}(A,B)]}{\sqrt{\mathrm{Var}[\overline{\Theta}_{\mathrm{u}}(A,B)]}} \geq \frac{\sqrt{k}}{1+\Upsilon(g,h_A)} e^{-\delta(a^*,b^*)}.$$

In comparison, choosing arbitrary leaves $a \in A$ and $b \in B$ and estimating $\overline{\Theta}_{\mathrm{u}}(A, B)$ using

$\widehat{\Theta}(a,b)$ one gets a signal-to-noise ratio of

$$\frac{\mathbf{E}[\widehat{\Theta}(a,b)]}{\sqrt{\mathrm{Var}[\widehat{\Theta}(a,b)]}} \;=\; \frac{\sqrt{k}}{1-e^{-2\delta(a,b)}}e^{-\delta(a,b)}.$$

Note that for large trees, $e^{-\delta(a,b)}$ is in general much smaller than $e^{-\delta(a^*,b^*)}$.

2.3 Idealized Example

The example of Figure 2A of the main text is defined more formally as follows. The example is made of three disjoint monophyletic subsets of leaves A, B, and C with corresponding subtrees $T_A = (V_A, E_A)$, $T_B = (V_B, E_B)$, and $T_C = (V_C, E_C)$ rooted respectively at a^*, b^*, and c^*. The triplet connecting the three subtrees is $q^* = a^*b^*|c^*$ and has distance matrix

$$\delta(a^*, b^*) = 2f_0,\ \delta(a^*, c^*) = \delta(b^*, c^*) = 4f_0,$$

where $f_0 > 0$. Refer to q^* as the "deep triplet." The subtrees T_B and T_C are complete binary trees with $H_0 + h_0$ levels and branch lengths $\delta(e) = g_0$ for all $e \in E_B \cup E_C$, with $g_0 > 0$. The subtree T_A is a complete binary tree with $H_0 + h_0$ levels and branch lengths g_0 modified in the following way: replace the subtree below the first node at level H_0 below a^* with a complete binary tree with h_1 levels and branch lengths $g_1 < g_0$. Refer to the latter subtree as the "dense subtree of T_A" and denote it by $T_d^* = (V_d^*, E_d^*)$. Denote by a_d^* the root of T_d^*. Under the molecular clock assumption, $h_1 g_1 = h_0 g_0$. In the current section, the full tree obtained this way is denoted by T.

The goal here is to infer the deep triplet, assuming that T_A, T_B, and T_C are known. More specifically, one seeks to analyze the performance of UPGMA and WPGMA on this example. Essentially, this comes down to analyzing the performance of the test

$$\bar{\delta}_{\mathrm{u}}(A, C) - \bar{\delta}_{\mathrm{u}}(A, B) > 0?$$

or, equivalently,

$$\overline{\Theta}_{\mathrm{u}}(A,C) - \overline{\Theta}_{\mathrm{u}}(A,B) < 0?$$

2.3.1 Variance formula

A general formula for the variance of the estimator

$$\widehat{\mathcal{L}} = \Theta_A^{-1}\Theta_B^{-1}(\overline{\Theta}_{\mathrm{u}}(A,C) - \overline{\Theta}_{\mathrm{u}}(A,B)),$$

is first derived. Note that $\overline{\Theta}_{\mathrm{u}}(A,C)$ and $\overline{\Theta}_{\mathrm{u}}(A,B)$ are not independent. One could derive a formula for the variance of $\widehat{\mathcal{L}}$ by using the full variance-covariance matrix of the distance matrix as computed in Section 1. Instead a "conditioning trick" is used. Noting that $\Theta_B = \Theta_C$,

$$\begin{aligned}
\widehat{\mathcal{L}} &= \Theta_A^{-1}\Theta_B^{-1}\left(\sum_{a\in A}\sum_{c\in C} w(a)w(c)\frac{1}{k}\sum_{i=1}^{k}\sigma_a^i\sigma_c^i - \sum_{a\in A}\sum_{b\in B} w(a)w(b)\frac{1}{k}\sum_{i=1}^{k}\sigma_a^i\sigma_b^i\right) \\
&= \frac{1}{k}\sum_{i=1}^{k}\left(\sum_{a\in A}\frac{w(a)\sigma_a^i}{\Theta_A}\right)\left(\sum_{c\in C}\frac{w(c)\sigma_c^i}{\Theta_C} - \sum_{b\in B}\frac{w(b)\sigma_b^i}{\Theta_B}\right) \\
&= \frac{1}{k}\sum_{i=1}^{k} S_A^i(S_C^i - S_B^i).
\end{aligned}$$

Let $\widehat{\mathcal{L}}_1 = S_A^1(S_C^1 - S_B^1)$. Let z^* be the meeting point of the deep triplet. Because of reversibility and the symmetries of the JC model, it suffices to perform the variance calculation conditioned on $\sigma_{z^*}^1 = 1$. In particular, this makes S_A^1, S_B^1, and S_C^1 independent. Denote the corresponding expectation with a star. Using formulas from (*S5*)

$$\begin{aligned}
\mathbf{E}^*[\widehat{\mathcal{L}}_1] &= \mathbf{E}^*[S_A^1](\mathbf{E}^*[S_C^1] - \mathbf{E}^*[S_B^1]) \\
&= e^{-f_0}(e^{-3f_0} - e^{-f_0}),
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{E}^*[\widehat{\mathcal{L}}_1^2] &= \mathbf{E}^*[(S_A^1)^2]\mathbf{E}^*\left[\left(S_C^1 - S_B^1\right)^2\right] \\
&= \mathbf{E}^*[(S_A^1)^2]\mathbf{E}^*\left[(S_C^1)^2 + (S_B^1)^2 - 2S_C^1 S_B^1\right] \\
&= \mathrm{Var}[S_A^1](\mathrm{Var}[S_C^1] + \mathrm{Var}[S_B^1] - 2\mathbf{E}^*[S_C^1]\mathbf{E}^*[S_B^1]) \\
&= 2(1 + K_{a^*,w})(1 + K_{b^*,w} - e^{-4f_0}).
\end{aligned}$$

Hence, one has (by symmetry)

$$\mathbf{E}[\widehat{\mathcal{L}}] = e^{-4f_0} - e^{-2f_0}, \tag{5}$$

and

$$\mathrm{Var}[\widehat{\mathcal{L}}] = \frac{1}{k}\left(2(1 + K_{a^*,w})(1 + K_{b^*,w} - e^{-4f_0}) - (e^{-4f_0} - e^{-2f_0})^2\right). \tag{6}$$

2.3.2 Gaussian approximation

By the Central Limit Theorem, quantities such as $\hat{\delta}_{\mathrm{u}}(a,b)$ and $\widehat{\Theta}(a,b)$ are well approximated by a Normal distribution. One can therefore obtain approximate formulas for the success probability if it is assumed that the distance matrix $(\hat{\delta}_{\mathrm{u}}(a,b))_{a,b\in[n]}$ is jointly Gaussian with variance-covariance matrix as computed in Section 1. In particular, the quantity $\widehat{\mathcal{L}}$ is Gaussian, as a linear combination of Gaussians, with expectation and variance as in Eqs. (5) and (6).

For instance, use the WPGMA weighting scheme in the case where T is such that $h_1 = h_0 = 0$, that is, there is no dense subtree. In that case, it was already computed $K_{a^*,w} = K_{b^*,w}$ in Eq. (4). One gets that for the test to fail with probability less than 1%, the sample complexity required is given by solving

$$\mathbf{E}[\widehat{\mathcal{L}}] \approx (2.3)\sqrt{\mathrm{Var}[\widehat{\mathcal{L}}]},$$

that is,

$$k \approx (2.3)^2\left(\frac{2(1 + \Upsilon(g_0,H_0))(1 + \Upsilon(g_0,H_0) - e^{-4f_0})}{(e^{-4f_0} - e^{-2f_0})^2} - 1\right). \tag{7}$$

As $f_0 \to 0$ and $H_0 \to +\infty$, one can check that this expression is roughly

$$k \approx \begin{cases} \frac{2(2.3)^2}{4} \cdot \left(\frac{(1-e^{-2g_0})e^{-2(g^{**}-g_0)}}{1-e^{-2(g^{**}-g_0)}} \right)^2 \cdot \frac{1}{f_0^2} & \text{if } g_0 < g^{**} \\ \frac{2(2.3)^2}{4} \cdot \left(\frac{1-e^{-2g_0}}{e^{2(g_0-g^{**})}-1} \right)^2 \cdot \frac{1}{f_0^2} \cdot e^{4(g_0-g^{**})H_0} & \text{if } g_0 \geq g^{**}, \end{cases}$$

consistent with Eq. (3) in the main text. In fact, this rough behavior holds for general topologies under the full Markov model, as shown in the attached paper. In Figure 2B of the main text, Eq. (7) is plotted as a function of g_0 for various values of H_0.

2.3.3 Lower bound for UPGMA

Although the use of UPGMA was justified through the independent-error model in Section 2.2, here the full Markov model is used to analyze the behavior of UPGMA. Let $h_1 > 0$ and $h_0 > 0$.

A lower bound on $K_{a^*,w}$ under the UPGMA weighting is given as follows. Recall that

$$K_{a^*,w} = \sum_{e \in E_A} R_{a^*}(e) w(e)^2,$$

with

$$R_{a^*}(e) = \left(1 - \theta_e^2\right) \Theta(a^*, y)^{-2},$$

where $e = (x, y)$ (x is assumed closer to a^*). Let K^{**} be the contribution of the edge immediately above the dense subtree, which is denote by e^{**}. That is,

$$K^{**} = R_{a^*}(e^{**}) w(e^{**})^2.$$

Clearly $K_{a^*,w} \geq K^{**}$. The fraction of leaves in the dense subtree is

$$\gamma = \frac{2^{h_1}}{2^{H_0+h_0} - 2^{h_0} + 2^{h_1}} = \frac{1}{2^{H_0+h_0-h_1} - 2^{h_0-h_1} + 1}.$$

and one gets

$$K^{**} = (1 - e^{-2g_0}) e^{2H_0 g_0} \gamma^2.$$

The signal-to-noise ratio of $\widehat{\mathcal{L}}$ is given by

$$\begin{aligned}\frac{|\mathbf{E}[\widehat{\mathcal{L}}]|}{\sqrt{\mathrm{Var}[\widehat{\mathcal{L}}]}} &= \frac{e^{-2f_0}-e^{-4f_0}}{\sqrt{\frac{1}{k}\left(2(1+K_{a^*,w})(1+K_{b^*,w}-e^{-4f_0})-(e^{-4f_0}-e^{-2f_0})^2\right)}} \\ &\leq \sqrt{\frac{k}{2e^{2H_0g_0}\gamma^2-1}}.\end{aligned}$$

If one chooses values of h_0, h_1 and H_0 such that $H_0 \geq h_0$ and such that the fraction of leaves in the dense subtree is comparable to the number of leaves in T_A, then γ is strictly positive (independently of H_0) and the signal-to-noise ratio goes to 0 unless k is exponential in H_0g_0 (roughly proportional to the depth). This exponential growth, which holds for any value of g_0, can be seen in Figure 3 of the main text where simulation results under the Jukes-Cantor model with the choice of parameters $h_0 = H_0$ and $h_1 = 1$ (giving $\gamma \geq 1/3$) are given. This particular example is somewhat extreme in that $g_1 = g_0/H_0$ is very small. Another example is given by $h_1 = H_0$ and $h_0 = 2H_0$ in which case $\gamma \geq 1/2$ and $g_1 = g_0/2$.

2.4 Probabilistic Analysis of WPGMA

Let $0 < f < g < +\infty$ and denote by $\mathbf{UY}^{f,g}$ the set of all phylogenies $\mathcal{T} = (V, E, [n], r; \delta) \in \mathbf{Y}^{f,g}$ where it holds further that δ is ultrametric, that is, for all $v \in V$ it holds that $\delta(v, x) = \delta(v, y) \equiv \delta(v)$, for all leaves x, y below v. Recall the WPGMA algorithm in Figure 1. WPGMA is run with the uncorrected distance estimates

$$\hat{\delta}_{\mathrm{u}}(a, b) = \frac{1}{2}(1 - \widehat{\Theta}(a, b)),$$

where

$$\widehat{\Theta}(a, b) = \nu^\top \widehat{F}^{ab}\nu = \frac{1}{k}\sum_{i=1}^{k}\sigma_a^i\sigma_b^i,$$

for $a, b \in [n]$. Call a subset of leaves A a clade if it corresponds to all leaf descendants of an internal node a^* called the MRCA. For a clade A with MRCA a^* and a leaf $a \in A$, let

$|a|_A = |a|_{a^*}$ and $\Theta_A = \Theta(a^*, a)$, where $|a|_{a^*}$ is the number of branches between a and the MRCA a^* of A. For disjoint clades A and B, let

$$\bar{\delta}_{\mathrm{u}}(A, B) = \sum_{a \in A} \sum_{b \in B} w(a) w(b) \hat{\delta}_{\mathrm{u}}(a, b) = \frac{1}{2}(1 - \overline{\Theta}_{\mathrm{u}}(A, B)),$$

where

$$\overline{\Theta}_{\mathrm{u}}(A, B) = \sum_{a \in A} \sum_{b \in B} w(a) w(b) \widehat{\Theta}(a, b),$$

and

$$w(a) = 2^{-|a|_A}.$$

Define

$$\Theta(a, b) = e^{-\delta(a,b)},$$

and similarly for $\Theta_{\mathrm{u}}(A, B)$.

The following theorem is proved in the Appendix.

Theorem 1 (Analysis of WPGMA)**.** For all $0 < f < g < g^{**}$, WPGMA solves the phylogenetic reconstruction problem on $\mathbf{UY}^{f,g} \otimes \{Q\}$ with $k = O(\log n)$.

3 General case

In this section, the molecular clock assumption is dropped and a general time-reversible matrix Q with stationary distribution $\pi > 0$ is considered. The following result is proved.

Definition 1 (Δ-Branch Model)**.** Let $0 < \Delta \leq f \leq g < +\infty$ and denote by $\mathbf{Y}_{\Delta}^{f,g}$ the set of all phylogenies $\mathcal{T} = (V, E, [n], r; \delta)$ satisfying $f \leq \delta(e) \leq g$ where $\delta(e)$ is an integer multiple of Δ, for all $e \in E$. Call $\mathbf{Y}_{\Delta}^{f,g} \otimes \{Q\}$ the Δ-Branch Model (Δ-BM).

Let $g^{**} = \ln \sqrt{2}$.

Theorem 2 (Main Result)**.** For all $0 < \Delta \leq f \leq g < g^{**}$, there is a distance-based method solving the phylogenetic reconstruction problem on $\mathbf{Y}_{\Delta}^{f,g} \otimes \{Q\}$ with $k = \kappa \log n$ for some constant $\kappa > 0$. As $\Delta \to 0$ (for fixed g), the constant κ scales as $O(\Delta^{-2})$.

A weaker version of the result stated here was first reported without proof in (*S6*). Note that in (*S6*) the result was stated without the discretization assumption which is in fact needed for the final step of the proof. This is further explained in Section 7.3 of (*S7*).

All proofs can be found in the Appendix.

3.1 Ancestral Reconstruction and Distance Averaging

Note that, without loss of generality, one can consider performing ancestral state reconstruction on a homogeneous tree as it is always possible to "complete" a general tree with zero-length edges.

Example 2 (Homogeneous Tree)**.** For an integer $h \geq 0$, denote by

$$\mathcal{T}^{(h)} = (V^{(h)}, E^{(h)}, L^{(h)}, r^{(h)}; \delta),$$

a rooted phylogeny where $T^{(h)}$ is the h-level complete binary tree with arbitrary edge weight function δ and $L^{(h)} = [2^h]$. For $0 \leq h' \leq h$, let $L_{h'}^{(h)}$ be the vertices on level $h - h'$ (from the root). In particular, $L_0^{(h)} = L^{(h)}$ and $L_h^{(h)} = \{r^{(h)}\}$.

In this section the discussion is hence restricted to the homogeneous case

$$\mathcal{T} = \mathcal{T}^{(h)} = (V, E, [n], r; \delta),$$

where $h = \log_2 n$, $f \leq \delta(e) \leq g$ and $\delta(e)$ is an integer multiple of Δ, $\forall e \in E$. Throughout this section, a sequence length $k > \kappa \log(n)$ is used where κ is a constant to be determined later. Generate k i.i.d. samples $(s_V^i)_{i=1}^k$ from the GTR model $(\mathcal{T}, Q)$ with state space $\Phi = \{\mathtt{A}, \mathtt{G}, \mathtt{C}, \mathtt{T}\}$.

3.2 Distance Estimator

Let the right eigenvector ν correspond to the second eigenvalue Λ_2 of Q. For $a, b \in [n]$, consider the estimator

$$\widehat{\Theta}(a,b) = \nu^\top \widehat{F}^{ab} \nu, \tag{8}$$

where $\widehat{F}^{ab}$ is the correlation matrix . For $a \in [n]$ and $i = 1, \ldots, k$, let

$$\sigma_a^i = \nu_{s_a^i}.$$

Then (8) is equivalent to

$$\widehat{\Theta}(a,b) = \frac{1}{k} \sum_{i=1}^{k} \sigma_a^i \sigma_b^i. \tag{9}$$

The next lemma indicates that this is indeed a legitimate distance estimator. In fact, $\widehat{\Theta}$ is a similarity estimator rather than a distance estimator. Similarity is used here for technical reasons. For more on connections between eigenvalues of the rate matrix and distance estimation, see e.g. (*S8, S9*).

Lemma 1 (Distance Estimator)**.** For all $a, b \in [n]$,

$$\mathbf{E}[\widehat{\Theta}(a,b)] = \Theta(a,b),$$

where $\Theta(a,b) = e^{-\delta(a,b)}$.

3.2.1 Ancestral Sequence Reconstruction

Let $e = (x,y) \in E$ and assume that x is closest to r (in topological distance). Define $\mathrm{Path}(r,e) = \mathrm{Path}(r,y)$, $|e|_r = |\mathrm{Path}(r,e)|$, and

$$R_r(e) = \left(1 - \theta_e^2\right) \Theta(r,y)^{-2},$$

where $\Theta(r,y) = e^{-\delta(r,y)}$ and $\theta_e = e^{-\delta(e)}$.

Proposition 1 below is a variant of Lemma 5.3 in (*S4*). For completeness, a proof is given.

Proposition 1 (Weighted Majority: GTR Version)**.** Let $s_{[n]}$ be a sample from the GTR model on $(\mathcal{T}, Q)$ with corresponding $\sigma_{[n]}$. For a unit flow w from r to $[n]$, consider the estimator

$$S = \sum_{x \in [n]} \frac{w(x)\sigma_x}{\Theta(r,x)}.$$

Then,

$$\mathbf{E}[S] = 0,$$

$$\mathbf{E}[S \,|\, \sigma_r] = \sigma_r,$$

and

$$\mathrm{Var}[S] = 1 + K_w,$$

where

$$K_w = \sum_{e \in E} R_r(e) w(e)^2.$$

Let w be a unit flow from r to $[n]$. The following multiplicative decomposition of w will be used: If $w(x) > 0$, let

$$\psi(e) = \frac{w(y)}{w(x)},$$

and, if instead $w(x) = 0$, let $\psi(y) = 0$. Denoting $x_\uparrow$ the immediate ancestor of $x \in V$ and letting $\theta_x = e^{-\delta((x_\uparrow, x))}$, it will be useful to re-write

$$K_w = \sum_{h'=0}^{h-1} \sum_{x \in L_{h'}^{(h)}} (1 - \theta_x^2) \prod_{e \in \mathrm{Path}(r,x)} \frac{\psi(e)^2}{\theta_e^2}, \tag{10}$$

and to define the following recursion from the leaves. For $x \in [n]$,

$$K_{x,w} = 0.$$

Then, let $u \in V - [n]$ with children v_1, v_2 with corresponding edges e_1, e_2 and define

$$K_{u,w} = \sum_{\alpha=1,2} ((1 - \theta_{v_\alpha}^2) + K_{v_\alpha,w}) \left(\frac{\psi(e_\alpha)^2}{\theta_{e_\alpha}^2} \right).$$

Note that, from (10), $K_{r,w} = K_w$.

Because short sequences are used, bounds on the variance are not enough: exponential concentration is needed. To obtain such concentration, the exponential moment of S is bounded. The proof generalizes a recent argument of Peres and Roch (arxiv.org/abs/0908.2056).

Proposition 2 (Weighted Majority: Exponential Bound)**.** For $\zeta \in \mathbb{R}$, let

$$\Gamma^i(\zeta) = \ln \mathbf{E}[\exp(\zeta S) \,|\, \sigma_r = \nu_i].$$

Then, there exists $c > 0$ depending only on Q and f such that for all $\zeta \in \mathbb{R}$,

$$\Gamma^i(\zeta) \leq \nu_i \zeta + \frac{1}{2} c \zeta^2 K_w.$$

3.2.2 Distance Averaging

The input to the tree reconstruction algorithm is the matrix of all estimated similarities between pairs of leaves $\{\widehat{\Theta}(a,b)\}_{a,b,\in[n]}$. For short sequences, these estimated similarities are known to be accurate for leaves that are close enough. It is now shown how to compute distances between internal nodes in a way that involves only $\{\widehat{\Theta}(a,b)\}_{a,b,\in[n]}$ (and previously computed internal weights) using Proposition 2.

Let $0 \leq h' < h$. For $v \in L_{h'}^{(h)}$, let $T_v = (V_v, E_v)$ be the subtree of $T = T^{(h)}$ rooted at v with leaf set denoted L_v. Let $a, b \in L_{h'}^{(h)}$. For $x \in \{a, b\}$, denote by X the leaves of $T = T^{(h)}$ below x. Assume that θ_e is given, for all e below a, b. Estimate $\delta(a,b)$ as follows

$$\bar{\delta}_{\mathrm{c}}(a,b) \;\equiv\; -\ln\left(\sum_{a' \in A}\sum_{b' \in B} w(a')w(b')\Theta(a,a')^{-1}\Theta(b,b')^{-1}\widehat{\Theta}(a',b')\right),$$

where

$$w(a') = \frac{1}{|A|}.$$

This choice of estimator is suggested by the following observation

$$\begin{aligned}\bar{\delta}_{\mathrm{c}}(a,b) &\equiv -\ln\left(\frac{1}{k}\sum_{i=1}^{k}\left(\sum_{a'\in A} w(a')\Theta(a,a')^{-1}\sigma^i_{a'}\right)\left(\sum_{b'\in B} w(b')\Theta(b,b')^{-1}\sigma^i_{b'}\right)\right)\\ &= -\ln\left(\frac{1}{k}\sum_{i=1}^{k}\left(\sum_{a'\in A}\frac{2^{-h'}\sigma^i_{a'}}{\Theta(a,a')}\right)\left(\sum_{b'\in B}\frac{2^{-h'}\sigma^i_{b'}}{\Theta(b,b')}\right)\right).\end{aligned}$$

Note that the first line depends only on estimates $(\widehat{\Theta}(u,v))_{u,v\in[n]}$ and $\{\Theta(v,\cdot)\}_{v\in V_a\cup V_b}$. The last line is the empirical distance between the reconstructed states at a and b when the flow is chosen to be homogeneous in Proposition 1.

Lemma 2 (Large Deviations)**.** Let $0 \le h' < h$ and let $a, b \in L^{(h)}_{h'}$. For $x = a, b$, let

$$S_x = \sum_{x'\in X}\frac{2^{-h'}\sigma_{x'}}{\Theta(x,x')}.$$

It holds that

$$-\ln\left(\mathbf{E}\left[\sum_{a'\in A}\sum_{b'\in B} w(a')w(b')\Theta(a,a')^{-1}\Theta(b,b')^{-1}\widehat{\Theta}(a',b')\right]\right) = \Theta(a,b),$$

where $\Theta(a,b) = e^{-\delta(a,b)}$ and there exists $\zeta^* > 0$ small enough such that

$$\mathbf{E}[\exp(\zeta S_a S_b)] < +\infty,$$

for all $|\zeta| < |\zeta^*|$. In particular, for all $\varepsilon > 0$ there exists $0 < \chi < 1$ such that

$$\mathbf{P}\left[\left|e^{-\bar{\delta}_{\mathrm{c}}(a,b)} - e^{-\delta(a,b)}\right| > \varepsilon\right] \le \chi^k.$$

Moreover, χ is a constant independent of h' and it scales as $1 - \Omega(\varepsilon^2)$ as $\varepsilon \to 0$.

In the next section, the previous lemma is used in two situations: 1) to estimate the distance between two close vertices; 2) to detect that two vertices are "far apart." These specializations of Lemma 2 are stated below.

Proposition 3 (Deep Distance Computation: Small Diameter)**.** Let $D > 0$, $\gamma > 0$, and $\varepsilon > 0$. Let $a, b \in L_{h'}^{(h)}$ as above. There exists $\kappa > 0$ such that if the following conditions hold:

- [Small Diameter] $\delta(a, b) < D$,
- [Sequence Length] $k > \kappa \log(n)$,

then

$$|\bar{\delta}_{\mathrm{c}}(a, b) - \delta(a, b)| < \varepsilon,$$

with probability at least $1 - O(n^{-\gamma})$. As $\varepsilon \to 0$, κ scales as $O(\varepsilon^{-2})$.

Proposition 4 (Deep Distance Computation: Diameter Test)**.** Let $D > 0$, $W > 5$, and $\gamma > 0$. Let $a, b \in L_{h'}^{(h)}$ as above. There exists $\kappa > 0$ such that if the following conditions hold:

- [Large Diameter] $\delta(a, b) > D + \ln W$,
- [Sequence Length] $k > \kappa \log(n)$,

then

$$\bar{\delta}_{\mathrm{c}}(a, b) > D + \ln \frac{W}{2},$$

with probability at least $1 - n^{-\gamma}$. On the other hand, if the first condition above is replaced by

- [Small Diameter] $\delta(a, b) < D + \ln \frac{W}{5}$,

then

$$\bar{\delta}_{\mathrm{c}}(a, b) \leq D + \ln \frac{W}{4},$$

with probability at least $1 - n^{-\gamma}$. The constant κ depends only on D and W.

3.3 Tree Reconstruction

In this section, the main result of this paper is proved. For $\Delta > 0$ and $z \in \mathbb{R}_+$, let $[z]_\Delta$ be the closest multiple of Δ to z (breaking ties arbitrarily).

3.3.1 Basic Definitions

An algorithm of (*S7*) called Blindfolded Cherry Picking is used. For reference, it is detailed in Figure 8. The reader is referred to (*S7*) for a full explanation of the algorithm, which is somewhat involved. The proof in (*S7*) is modular and relies on two main components: a distance-based combinatorial argument which remains unchanged in the setting here; and a statistical argument which is now adapted. The key to the latter is (*S7, Proposition 4*). Note that (*S7, Proposition 4*) is not distance-based as it relies on a complex ancestral reconstruction function—recursive majority. The main contribution in this section is to show how this result can be obtained using the techniques of the previous sections—leading to a fully distance-based reconstruction algorithm.

A few definitions from (*S7*) are needed. The interested reader is strongly advised to consult (*S7*) for a full explanation and motivation of these definitions.

Fix $0 < \Delta \leq f \leq g < g^{**}$ as in Theorem 2. Let $\mathcal{T} = (V, E, [n], r; \delta) \in \mathbf{Y}_{\Delta}^{f,g}$ be a phylogeny with underlying tree $T = (V, E)$. In this section, the edge set, vertex set and leaf set of a tree T' are sometimes referred to as $\mathcal{E}(T')$, $\mathcal{V}(T')$, and $\mathcal{L}(T')$ respectively.

Definition 2 (Restricted Subtree)**.** Let $V' \subseteq V$ be a subset of the vertices of T. The subtree of T restricted to V' is the tree T' obtained by 1) keeping only nodes and edges on paths between vertices in V' and 2) by then contracting all paths composed of vertices of degree 2, except the nodes in V'. The notation $T' = T|_{V'}$ is sometimes used. See Figure 2 for an example.

Definition 3 (Edge Disjointness)**.** Denote by $\mathrm{Path}_T(x, y)$ the path (sequence of edges) connecting x to y in T. Say that two restricted subtrees T_1, T_2 of T are edge disjoint if

$$\mathrm{Path}_T(x_1, y_1) \cap \mathrm{Path}_T(x_2, y_2) = \emptyset,$$

for all $x_1, y_1 \in \mathcal{L}(T_1)$ and $x_2, y_2 \in \mathcal{L}(T_2)$. Say that T_1, T_2 are edge sharing if they are not edge disjoint. See Figure 3 for an example.

Definition 4 (Legal Subforest)**.** Say that a tree is a rooted full binary tree if all its internal nodes have degree 3 except the root which has degree 2. A restricted subtree T_1 of T is a legal subtree of T if it is also a rooted full binary tree. Say that a forest

$$\mathcal{F} = \{T_1, T_2, \ldots\},$$

is legal subforest of T if the T_ι's are edge-disjoint legal subtrees of T. Denote by $\rho(\mathcal{F})$ the set of roots of $\mathcal{F}$.

Definition 5 (Dangling Subtrees)**.** Say that two edge-disjoint legal subtrees T_1, T_2 of T are dangling if there is a choice of root for T not in T_1 or T_2 that is consistent with the rooting of both T_1 and T_2. See Figure 4 below for an example where two legal, edge-disjoint subtrees are not dangling.

Definition 6 (Basic Disjoint Setup (General))**.** Let $T_1 = T_{x_1}$ and $T_2 = T_{x_2}$ be two restricted subtrees of T rooted at x_1 and x_2 respectively. Assume further that T_1 and T_2 are edge-disjoint, but not necessarily dangling. Denote by y_ι, z_ι the children of x_ι in T_ι, $\iota = 1, 2$. Let w_ι be the node in T where the path between T_1 and T_2 meets T_ι, $\iota = 1, 2$. Note that w_ι may not be in T_ι since T_ι is restricted, $\iota = 1, 2$. If $w_\iota \neq x_\iota$, assume without loss of generality that w_ι is in the subtree of T rooted at z_ι, $\iota = 1, 2$. Call this configuration the Basic Disjoint Setup (General). See Figure 4. Let $\delta(T_1, T_2)$ be the length of the path between w_1 and w_2 in the metric δ.

Example 3. Consider the example of Figure 5. The tree A is a complete binary tree with H_0+h_0 levels and branch lengths g_0 with one subtree at height H_0 replaced by a complete binary tree with h_1 levels (dark subtree) which is denser, that is, with branch lengths $g_1 < g_0$. Assume to simplify that the molecular clock assumption is satisfied, that is, $h_1 g_1 = h_0 g_0$. Imagine that the first h levels of A where $h = h_0 + 2 < h_1$ have been reconstructed. Then a rooted forest composed of 2^{h_1-h} subtrees of the dense subtree and $2^{H_0+h_0-h}$ subtrees of the sparse part of the

tree is obtained. Denote by B and C the leftmost dense and sparse subtree respectively. The relation between these two subtrees is illustrated in Figure 6. The configuration obtained satisfies the conditions of the Basic Disjoint Setup (General) but is not dangling: the two subtrees are disjoint but the path connecting them does not enter through their respective roots.

3.3.2 Deep Distorted Metric

The goal in this subsection is to compute the distance between the internal nodes x_1 and x_2 in the Basic Disjoint Setup (General). It has already been shown how to perform this computation when T_1 and T_2 are dangling, as this case is handled easily by Propositions 3 and 4. However, in the general case depicted in Figure 4, there is a complication. When T_1 and T_2 are not dangling, the reconstructed sequences at x_1 and x_2 are not conditionally independent. But it can be shown that for the algorithm Blindfolded Cherry Picking to work properly, the following are needed: 1) to compute the distance between x_1 and x_2 correctly when the two subtrees are close and dangling; 2) to detect when the two subtrees are far apart (but an accurate distance estimate is not required in that case). This turns out to be enough because the algorithm Blindfolded Cherry Picking ensures roughly that close reconstructed subtrees are always dangling. The reader is referred to (*S7*) for details.

The key point is the following: if one computes the distance between y_1 and y_2 rather than the distance between x_1 and x_2, then the dangling assumption is satisfied (re-root the tree at any node along the path connecting w_1 and w_2). However, when the algorithm has only reconstructed T_1 and T_2, one cannot tell which pair in $\{y_1, z_1\} \times \{y_2, z_2\}$ is the right one to use for the distance estimation. Instead, compute the distance for all pairs in $\{y_1, z_1\} \times \{y_2, z_2\}$ and the following then holds: in the dangling case, all these distances will agree (after subtracting the length of the edges between x_1, x_2 and $\{y_1, z_1, y_2, z_2\}$); in the general case, at least one is correct. This is the basic observation behind the routine DISTORTEDMETRIC in Figure 7 and

the proof of Proposition 5 below.

Using the notation of Definition 6, fix $(a,b) \in \{y_1, z_1\} \times \{y_2, z_2\}$. For $x = a, b$, denote by X the leaves of T_x and let $|\ell|_x$ be the graph distance (that is, the number of edges) between x and leaf $\ell \in X$. Assume that θ_e is given for all $e \in \mathcal{E}(T_a) \cup \mathcal{E}(T_b)$. The quantity $\delta(a,b)$ is estimated as follows

$$\bar{\delta}_{\mathrm{c}}(a,b) \;\equiv\; -\ln\left(\sum_{a'\in A}\sum_{b'\in B} w(a')w(b')\Theta(a,a')^{-1}\Theta(b,b')^{-1}\widehat{\Theta}(a',b')\right),$$

where

$$w(a') = 2^{-|a'|_a}.$$

Note that, because the tree is binary, it holds that

$$\sum_{a'\in A}\sum_{b'\in B} 2^{-|a'|_a-|b'|_b} = \sum_{a'\in A} 2^{-|a'|_a}\sum_{b'\in B} 2^{-|b'|_b} = 1,$$

and think of the weights on A (similarly for B) as resulting from a homogeneous flow w_a from a to A. Then, the bounds on the variance and the exponential moment of

$$S_a \equiv \sum_{a'\in A} 2^{-|a'|_a}\Theta(a,a')^{-1}\sigma_{a'},$$

in Propositions 1 and 2 still hold with

$$K_{a,w_a} = \sum_{e\in\mathcal{E}(T_a)} R_a(e)w(e)^2.$$

Moreover K_{a,w_a} is uniformly bounded following an argument identical to (4) in the proof of Lemma 2. In particular, the same large deviations result hold for $\bar{\delta}_{\mathrm{c}}(a,b)$.

For $D > 0$, $W > 5$, define

$$\overline{\mathbf{SD}}(a,b) = \mathbf{1}\left\{[\bar{\delta}_{\mathrm{c}}(a,b)]_\Delta \le D + \ln\frac{W}{3}\right\},$$

and let

$$\bar{\mathrm{d}}_{\mathrm{c}}(a,b) = \begin{cases} [\bar{\delta}_{\mathrm{c}}(a,b)]_\Delta, & \text{if } \overline{\mathbf{SD}}(a,b) = 1, \\ +\infty, & \text{o.w.} \end{cases}$$

Proposition 5 (Accuracy of DISTORTEDMETRIC)**.** Let $D > 0$, $W > 5$, $\gamma > 0$ and $g < g^{**}$. Consider the Basic Disjoint Setup (General) with $\mathcal{F} = \{T_1, T_2\}$ and $\mathcal{Q} = \{y_1, z_1, y_2, z_2\}$. Assume θ_e is given for all $e \in \mathcal{E}(T_1) \cup \mathcal{E}(T_2)$. Let Υ denote the output of DISTORTEDMETRIC in Figure 7. There exists $\kappa > 0$, such that if the following condition holds:

- [Edge Length] It holds that $\delta(e) \leq g$, $\forall e \in \mathcal{E}(T_x)$, $x \in \mathcal{Q}$;
- [Sequence Length] The sequence length is $k > \kappa \log(n)$,

then, with probability at least $1 - O(n^{-\gamma})$,

$$\Upsilon = \delta(x_1, x_2)$$

under either of the following two conditions:

1. [Dangling Case] T_1 and T_2 are dangling and $\delta(T_1, T_2) < D$, or
2. [Finite Estimate] $\Upsilon < +\infty$.

As $\Delta \to 0$ (for fixed D, W), the constant κ scales as $O(\Delta^{-2})$.

The rest of the Blindfolded Cherry Picking algorithm is unchanged. The full algorithm is given in Figure 8. The algorithm is quite complex and a full explanation is provided in (*S7*).

The description of the algorithm uses the following notation. Let $\varepsilon > 0$ be the error tolerance. Each application of Proposition 5 has an error of $O(n^{-\gamma})$. Note that $O(n^3)$ distance estimations are performed so that, by the union bound, one requires $O(n^{3-\gamma}) \leq \varepsilon$. Set $D > 10g$ and take W such that $\ln \frac{W}{3} > 4g$. For a distance matrix $\mathcal{D}$ and a set of nodes $\mathcal{N}$, denote

$$\overline{\mathcal{D}}(\mathcal{N}) = \max \{\mathcal{D}(x, y) : \{x, y\} \subseteq \mathcal{N}\}.$$

This concludes the sketch of the proof of Theorem 2.

4 The Distance Matrix is Not Sufficient

A statistic (i.e., a function of the full data) is called sufficient if, conditioned on the value of the statistic, the distribution of the full data does not depend on the parameters of the generating model. Roughly speaking, a sufficient statistic encapsulates all the information about the data. See e.g. (*S10*). In this section, it is shown that the pairwise correlation matrices do not constitute a sufficient statistic for the full Markov model of evolution. Hence, there is in principle more information in the full sequence dataset than there is in the matrix of evolutionary distances.

A simple example of non-sufficiency follows. Consider a four-leaf tree with leaf set $L = \{a, b, c, d\}$ and split $ab|cd$. Assume a CFN model with purines denoted "0" and pyrimidines denoted "1" with equal mutation probabilities p is used.

Example 4 (CFN Model)**.** The *CFN model* is the GTR model with $\phi = 2$, $\pi = (1/2, 1/2)$, and

$$Q = Q^{\mathrm{CFN}} \equiv \begin{pmatrix} -1/2 & 1/2 \\ 1/2 & -1/2 \end{pmatrix}.$$

Consider the following correlation matrices

$$\widehat{F}^{ij}_{\upsilon_1\upsilon_2} = \frac{1}{4},$$

for all $i \neq j \in L$ and $\upsilon_1, \upsilon_2 \in \{0, 1\}$. Two different datasets consistent with these correlation matrices are

$$\mathrm{Data}_1 = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix},$$

and

$$\mathrm{Data}_2 = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix},$$

where the columns are the sites and the rows are the leaves in the order a, b, c, d.

Compare the probability of observing the two datasets under two different values of p: $p = \varepsilon$ and $= 1/2 - \varepsilon$ for $\varepsilon > 0$ small. In the first case, in a first approximation it suffices to compute the parsimony scores and

$$\mathbf{P}_\varepsilon[\mathrm{Data}_1] = \left(\frac{\varepsilon}{2}\right)^8 + O(\varepsilon^9) = \frac{\varepsilon^8}{256} + O(\varepsilon^9),$$

and

$$\mathbf{P}_\varepsilon[\mathrm{Data}_2] = \left(\frac{1}{2}\right)^2 \left(\frac{\varepsilon}{2}\right)^2 \left(\varepsilon^2\right)^4 + O(\varepsilon^{11}) = \frac{\varepsilon^{10}}{16} + O(\varepsilon^{11}).$$

In particular, the following ratio is obtained

$$\frac{\mathbf{P}_\varepsilon[\mathrm{Data}_2 \,|\, \widehat{F}]}{\mathbf{P}_\varepsilon[\mathrm{Data}_1 \,|\, \widehat{F}]} = \frac{\mathbf{P}_\varepsilon[\mathrm{Data}_2]}{\mathbf{P}_\varepsilon[\mathrm{Data}_1]} = \varepsilon^2 + O(\varepsilon^3).$$

On the other hand, if $p = 1/2 - \varepsilon$ then the state distribution is almost uniform and

$$\frac{\mathbf{P}_{1/2-\varepsilon}[\mathrm{Data}_2 \,|\, \widehat{F}]}{\mathbf{P}_{1/2-\varepsilon}[\mathrm{Data}_1 \,|\, \widehat{F}]} = \frac{\mathbf{P}_{1/2-\varepsilon}[\mathrm{Data}_2]}{\mathbf{P}_{1/2-\varepsilon}[\mathrm{Data}_1]} = 1 + O(\varepsilon).$$

Since the ratios are different, it has been shown that the distribution of the data conditioned on the correlation matrices depends on the parameters of the model. Therefore, the distance matrix is not a sufficient statistic.

A Proofs

A.1 Section 2

Proof of Theorem 1: Fix $\overline{D} > 3g + 2f$, $2g + 2f < \underline{D} < \overline{D}$, and

$$\varepsilon' < \min\left\{\frac{e^{2f}-1}{e^{2f}+1}, \frac{e^{\underline{D}-2g-2f}-1}{e^{\underline{D}-2g-2f}+1}\right\}.$$

This choice ensures that

$$e^{2f}\frac{1-\varepsilon'}{1+\varepsilon'} > 1,$$

and

$$e^{\underline{D}-2g-2f}\frac{1-\varepsilon'}{1+\varepsilon'} > 1,$$

which will be needed later. Let

$$\varepsilon = \min\{\varepsilon' e^{-\overline{D}}, \varepsilon' e^{-\underline{D}}\},$$

and let χ be as in Lemma 2 in Section 3 for this choice of ε. Taking κ large enough, assume the conclusion of Lemma 2 holds for all pairs of clades in the tree, an event denoted by $(\star)$.

By definition,

$$\bar{\delta}_{\mathrm{u}}(A,B) \le \bar{\delta}_{\mathrm{u}}(A',B') \quad \Longleftrightarrow \quad \overline{\Theta}_{\mathrm{u}}(A,B) \ge \overline{\Theta}_{\mathrm{u}}(A',B').$$

For convenience, in the rest of the proof $\overline{\Theta}_{\mathrm{u}}$ is used rather than $\bar{\delta}_{\mathrm{u}}$. If A, B are disjoint clades with respective MRCA a^* and b^* satisfying $\delta(a^*, b^*) < \overline{D}$,

$$\begin{aligned}
\overline{\Theta}_{\mathrm{u}}(A,B) &< \Theta_{\mathrm{u}}(A,B) + \Theta_A\Theta_B\varepsilon \\
&\le \Theta_A\Theta_B(e^{-\delta(a^*,b^*)} + \varepsilon' e^{-\overline{D}}) \\
&< \Theta_A\Theta_B(e^{-\delta(a^*,b^*)} + \varepsilon' e^{-\delta(a^*,b^*)}) \\
&= \Theta_{\mathrm{u}}(A,B)(1+\varepsilon'),
\end{aligned}$$

and similarly

$$\overline{\Theta}_{\mathrm{u}}(A,B) \;>\; \Theta_{\mathrm{u}}(A,B)(1-\varepsilon').$$

On the other hand, if $\delta(a^*,b^*) > \underline{D}$,

$$\begin{aligned}
\overline{\Theta}_{\mathrm{u}}(A,B) &< \Theta_{\mathrm{u}}(A,B) + \Theta_A\Theta_B\varepsilon \\
&\leq \Theta_A\Theta_B(e^{-\delta(a^*,b^*)} + \varepsilon' e^{-\underline{D}}) \\
&< \Theta_A\Theta_B(e^{-\underline{D}} + \varepsilon' e^{-\underline{D}}) \\
&= \Theta_A\Theta_B e^{-\underline{D}}(1+\varepsilon').
\end{aligned}$$

By $(\star)$ these inequalities hold for all such pairs of clades.

Two clades A, B are sister clades if their MRCA is their immediate ancestor. The following convention is used. Recall that the leaves are denoted $\{1,\ldots,n\}$. Let $\min A$ be the smallest label in A. When denoting a pair of sister clades (A,B), assume $\min A < \min B$. There are $n-1$ pairs of sister clades. Order the sister pairs by decreasing value of $\overline{\Theta}_{\mathrm{u}}(A,B)$, breaking ties by lexicographic order over $(\min A, \min B)$:

$$(A_1,B_1),\ldots,(A_{n-1},B_{n-1}).$$

Assume that WPGMA uses the same tie-breaking rule. Let $C_i = A_i \cup B_i$.

The following basic claim is proved next. For all $i = 1,\ldots,n-1$, at Selection Step i choose $(A^*,B^*) = (A_i,B_i)$. The result then follows. The argument works by induction. For $i = 0$, there is nothing to prove. Assume the claim holds up to some $1 \leq i < n-1$. Observations:

1. All the current clusters in $\mathcal{Z}_{i-1}$ are clades. This follows from the induction hypothesis. By the induction hypothesis, one also gets that the values $\bar{\delta}_{\mathrm{u}}(A,B)$ computed at the Reduction Steps indeed correspond to the original definition:

$$\bar{\delta}_{\mathrm{u}}(A,B) = \sum_{a\in A}\sum_{b\in B} 2^{-|a|_A}2^{-|b|_B}\hat{\delta}_{\mathrm{u}}(a,b) = \frac{1-\overline{\Theta}_{\mathrm{u}}(A,B)}{2},$$

where

$$\overline{\Theta}_{\mathrm{u}}(A,B) = \sum_{a\in A}\sum_{b\in B} 2^{-|a|_A} 2^{-|b|_B} \widehat{\Theta}(a,b).$$

2. It is shown next that for all $C \in \mathcal{Z}_{i-1}$,

$$\Theta_{\mathrm{u}}(A_i,B_i)e^{-2f} < \Theta_C^2 \leq \Theta_{\mathrm{u}}(A_i,B_i)e^{2g+2f}.$$

Let $C \in \mathcal{Z}_{i-1}$ such that $C = A \cup B$ for sister clades A, B. By $(\star)$,

$$\begin{aligned}
\Theta_C^2 &= \Theta_{\mathrm{u}}(A,B) \\
&> \overline{\Theta}_{\mathrm{u}}(A,B)(1+\varepsilon')^{-1} \\
&> \overline{\Theta}_{\mathrm{u}}(A_i,B_i)(1+\varepsilon')^{-1} \\
&> \Theta_{\mathrm{u}}(A_i,B_i)\frac{1-\varepsilon'}{1+\varepsilon'} \\
&> \Theta_{\mathrm{u}}(A_i,B_i)e^{-2f}.
\end{aligned}$$

Conversely, if a clade $C = A \cup B$ with sister clades A, B satisfies

$$\Theta_C^2 = \Theta_{\mathrm{u}}(A,B) > \Theta_{\mathrm{u}}(A_i,B_i)e^{2f}, \tag{11}$$

then

$$\begin{aligned}
\overline{\Theta}_{\mathrm{u}}(A,B) &> (1-\varepsilon')\Theta_{\mathrm{u}}(A,B) \\
&> (1-\varepsilon')\Theta_{\mathrm{u}}(A_i,B_i)e^{2f} \\
&> (1-\varepsilon')\Theta_{\mathrm{u}}(A_i,B_i)\frac{1+\varepsilon'}{1-\varepsilon'} \\
&> (1+\varepsilon')\Theta_{\mathrm{u}}(A_i,B_i) \\
&> \overline{\Theta}_{\mathrm{u}}(A_i,B_i),
\end{aligned} \tag{12}$$

so that C must be included in a cluster of $\mathcal{Z}_i$ by the induction hypothesis. In particular, if two sister clades A, B are such that $\Theta_A^2, \Theta_B^2 > \Theta_{\mathrm{u}}(A_i,B_i)e^{2g+2f}$ then (11) is satisfied, that

is, $\Theta_{\mathrm{u}}(A,B) > \Theta_{\mathrm{u}}(A_i,B_i)e^{2f}$. By (12), (A,B) would have been selected in a previous iteration by induction. That implies, for all $C \in \mathcal{Z}_{i-1}$,

$$\Theta_C^2 \leq \Theta_{\mathrm{u}}(A_i,B_i)e^{2g+2f}.$$

3. Claim: $A_i, B_i \in \mathcal{Z}_{i-1}$. Indeed, by the previous paragraph all clades with Θ^2-value at least $\Theta_{\mathrm{u}}(A_i,B_i)e^{2f}$ have been constructed in a previous iteration. In particular, the clade A_i has been constructed in a previous step as it satisfies

$$\Theta_{A_i}e^{-f} > \Theta_{C_i} = \sqrt{\Theta_{\mathrm{u}}(A_i,B_i)}.$$

   The same holds for B_i. Moreover, A_i and B_i being sister clades of each other (and no other clades), they cannot have been selected inside another pair by the induction hypothesis.

4. By construction, (A_i,B_i) is chosen over all other sister clades present in $\mathcal{Z}_{i-1}$. So it remains to show that (A_i,B_i) is selected over all other pairs. Pairs of clades that are far enough will not be selected. That is, if A,B with MRCA a^*,b^* is such that

$$\delta(a^*,b^*) \geq \underline{D},$$

   then

$$\begin{aligned}
\overline{\Theta}_{\mathrm{u}}(A,B) &< \Theta_A\Theta_B e^{-\underline{D}}(1+\varepsilon') \\
&< \Theta_{\mathrm{u}}(A_i,B_i)e^{2g+2f}e^{-\underline{D}}(1+\varepsilon') \\
&< \overline{\Theta}_{\mathrm{u}}(A_i,B_i)(1-\varepsilon')^{-1}e^{2g+2f}e^{-\underline{D}}(1+\varepsilon') \\
&< \overline{\Theta}_{\mathrm{u}}(A_i,B_i),
\end{aligned}$$

   by assumption on ε'.

5. Finally, non-sister clades that are closer than $\underline{D}$ cannot be selected. Indeed, assume by contradiction that (A^*, B^*) is such a pair. Since (A^*, B^*) are not sister clades, at least one of them, say A^* without loss of generality, has an immediate ancestor u that is strictly lower than the MRCA of A^* and B^*. Take C^* to be any clade in $\mathcal{Z}_{i-1}$ below u that is different than A^*. There must be such a clade because otherwise A^* would have been merged with its sister already. The MRCA of A^* and C^* is u. Moreover, one must have

$$\Theta^2_{A^*} > \Theta_{\mathrm{u}}(A_i, B_i)e^{-2f},$$

and

$$\Theta^2_{C^*} \leq \Theta_{\mathrm{u}}(A_i, B_i)e^{2g+2f},$$

so that

$$\delta(a^*, c^*) < 2g + g + 2f < 3g + 2f < \overline{D},$$

where a^* and c^* are the MRCA of A^* and C^* respectively. Finally by $(\star)$

$$\begin{aligned}
\overline{\Theta}_{\mathrm{u}}(A^*, C^*) &> \Theta_{\mathrm{u}}(A^*, C^*)(1 - \varepsilon') \\
&> \Theta_{\mathrm{u}}(A^*, B^*)e^{2f}(1 - \varepsilon') \\
&> \overline{\Theta}_{\mathrm{u}}(A^*, B^*)(1 + \varepsilon')^{-1}e^{2f}(1 - \varepsilon') \\
&> \overline{\Theta}_{\mathrm{u}}(A^*, B^*).
\end{aligned}$$

This is a contradiction.

■

A.2 Section 3

Proof of Lemma 1: Note that $\mathbf{E}[\widehat{F}_{ij}^{ab}] = \pi_i \left(e^{-\delta(a,b)Q}\right)_{ij}$. Then

$$\begin{aligned}
\mathbf{E}\left[\nu^\top \widehat{F}^{ab}\nu\right] &= \sum_{i\in\Phi}\nu_i \sum_{j\in\Phi}\pi_i \left(e^{-\delta(a,b)Q}\right)_{ij}\nu_j \\
&= \sum_{i\in\Phi}\nu_i(\pi_i e^{-\delta(a,b)}\nu_i) \\
&= e^{-\delta(a,b)}\sum_{i\in\Phi}\pi_i\nu_i^2 \\
&= e^{-\delta(a,b)}.
\end{aligned}$$

■

Proof of Proposition 1: The proofs of (*S5, S4*) are followed below. Let $\bar{e}_i$ be the unit vector in direction i. Let $x \in [n]$, then

$$\mathbf{E}[\bar{e}_{s_x}^\top \mid s_r] = \bar{e}_{s_r}^\top e^{\delta(r,x)Q}.$$

Therefore,

$$\mathbf{E}[\sigma_x \mid \sigma_r] = \bar{e}_{s_r}^\top e^{\delta(r,x)Q}\nu = \sigma_r e^{-\delta(r,x)},$$

and

$$\mathbf{E}[S \mid \sigma_r] = \sum_{x\in[n]} \frac{w(x)\sigma_r e^{-\delta(r,x)}}{\Theta(r,x)} = \sigma_r \sum_{x\in[n]} w(x) = \sigma_r.$$

In particular,

$$\mathbf{E}[S] = \sum_{\iota\in\Phi}\pi_i\nu_i = 0.$$

For $x, y \in [n]$, let $x \wedge y$ be the meeting point of the paths between r, x, y. Note

$$\begin{aligned}
\mathbf{E}[\sigma_x \sigma_y] &= \sum_{\iota \in \Phi} \mathbf{P}[s_{x\wedge y} = \iota] \mathbf{E}[\sigma_x \sigma_y \mid s_{x\wedge y} = \iota] \\
&= \sum_{\iota \in \Phi} \pi_\iota \mathbf{E}[\sigma_x \mid s_{x\wedge y} = \iota] \mathbf{E}[\sigma_y \mid s_{x\wedge y} = \iota] \\
&= \sum_{\iota \in \Phi} \pi_\iota e^{-\delta(x\wedge y, x)} \nu_\iota e^{-\delta(x\wedge y, y)} \nu_\iota \\
&= e^{-\delta(x,y)} \sum_{\iota \in \Phi} \pi_\iota \nu_\iota^2 \\
&= e^{-\delta(x,y)}.
\end{aligned}$$

Then

$$\begin{aligned}
\mathrm{Var}[S] &= \mathbf{E}[S^2] \\
&= \sum_{x,y \in [n]} \frac{w(x)w(y)}{\Theta(r,x)\Theta(r,y)} \mathbf{E}[\sigma_x \sigma_y] \\
&= \sum_{x,y \in [n]} w(x)w(y) e^{2\delta(r, x\wedge y)}.
\end{aligned}$$

For $e \in E$, let $e = (e_\uparrow, e_\downarrow)$ where $e_\uparrow$ is the vertex closest to r. Then, by a telescoping sum, for $u \in V$

$$\begin{aligned}
\sum_{e \in \mathrm{Path}(r,u)} R_r(e) &= \sum_{e \in \mathrm{Path}(r,u)} e^{2\delta(r, e_\downarrow)} - \sum_{e \in \mathrm{Path}(r,u)} e^{2\delta(r, e_\uparrow)} \\
&= e^{2\delta(r,u)} - 1,
\end{aligned}$$

and therefore

$$\begin{aligned}
\mathbf{E}[S^2] &= \sum_{x,y \in [n]} w(x)w(y) e^{2\delta(v, x\wedge y)} \\
&= \sum_{x,y \in [n]} w(x)w(y) \left(1 + \sum_{e \in \mathrm{Path}(r, x\wedge y)} R_r(e)\right) \\
&= 1 + \sum_{e \in E} R_r(e) \sum_{x,y \in [n]} \mathbf{1}\{e \in \mathrm{Path}(r, x \wedge y)\} w(x)w(y) \\
&= 1 + \sum_{e \in E} R_r(e) w(e)^2.
\end{aligned}$$

■

Proof of Proposition 2: The claim is proved by induction, moving away from the leaves. The following analytical lemma is inspired by the proof of Peres and Roch.

Lemma 3 (Recursion Step)**.** Let $M = e^{\delta Q}$ with second right eigenvector ν and corresponding eigenvalue $\lambda = e^{-\delta}$ satisfying $\delta \geq f$. Then there is $c > 0$ depending on Q and f such that for all $i \in \Phi$

$$F(x) \equiv \sum_{j \in \Phi} M_{ij} \exp(\nu_j x) \leq \exp(\lambda \nu_i x + \frac{1}{2} c(1 - \lambda^2) x^2) \equiv G(x), \tag{13}$$

for all $x \in \mathbb{R}$.

Proof of Lemma 3: Let $c' = c(1 - \lambda^2)$. Note that

$$F'(x) = \sum_{j \in \Phi} M_{ij} \nu_j \exp(\nu_j x),$$

$$F''(x) = \sum_{j \in \Phi} M_{ij} \nu_j^2 \exp(\nu_j x),$$

$$G'(x) = (\lambda \nu_i + c' x) \exp(\lambda \nu_i x + \frac{1}{2} c' x^2),$$

and

$$G''(x) = ((\lambda \nu_i + c' x)^2 + c') \exp(\lambda \nu_i x + \frac{1}{2} c' x^2).$$

Hence,

$$F(0) = G(0) = 1,$$

$$F'(0) = G'(0) = \lambda \nu_i.$$

Let

$$\bar{\pi} = \min_{\iota} \pi_\iota,$$

and

$$\bar{\nu} \equiv \max_i |\nu_i| \leq \frac{1}{\sqrt{\bar{\pi}}}.$$

Note that

$$F''(x) \leq \bar{\nu}^2 \exp(\bar{\nu}|x|) \equiv \overline{F}(x),$$

and

$$G''(x) \geq c' \exp(-\bar{\nu}|x| + \frac{1}{2}c'x^2) \equiv \overline{G}(x).$$

Choose $c' = c^* > 0$ such that $\overline{F}(x) < \overline{G}(x)$ for all $x \in \mathbb{R}$. Note in particular that taking

$$c^* > \max\left\{4\bar{\nu}, \bar{\nu}^2 \exp(2\bar{\nu})\right\},$$

is enough. Indeed, for $|x| > 1$ it holds that $c^* > \bar{\nu}^2$ and $\exp(-\bar{\nu}|x| + \frac{1}{2}c^*x^2) > \exp(\bar{\nu}|x|)$ so that $\overline{F}(x) < \overline{G}(x)$. For $|x| \leq 1$,

$$\overline{G}(x) > c^* \exp(-\bar{\nu}) > \bar{\nu}^2 \exp(\bar{\nu}) \geq \overline{F}(x).$$

Now choose $c = c^*(1 - e^{-2f})^{-1}$ in (13) (which implies $c' \geq c^*$ by $\delta \geq f$). Then,

$$G''(x) \geq \overline{G}(x) > \overline{F}(x) \geq F''(x),$$

and therefore

$$G(x) \geq F(x),$$

for all $x \in \mathbb{R}$. ■

Going back to the proof of Proposition 2, let $S_x = \sigma_x$ for all $x \in [n]$ and

$$S_u = \sum_{\alpha=1,2} S_{v_\alpha} \frac{\psi(e_\alpha)}{\theta_{e_\alpha}},$$

where $u \in V - [n]$ with children v_1, v_2 with corresponding edges e_1, e_2. Note that $S_r = S$. Let

$$\Gamma_u^i(\zeta) = \ln \mathbf{E}[\exp(\zeta S_u) \,|\, \sigma_u = \nu_i].$$

Take $c > 0$ as in Lemma 3. The main claim is clearly true at the leaves, that is, for all $x \in [n]$

$$\begin{aligned}
\Gamma_x^i(\zeta) &= \ln \mathbf{E}[\exp(\zeta S_x) \,|\, \sigma_x = \nu_i] \\
&= \ln \mathbf{E}[\exp(\zeta \sigma_x) \,|\, \sigma_x = \nu_i] \\
&= \nu_i \zeta \\
&\leq \nu_i \zeta + \frac{1}{2} c \zeta^2 K_{x,w}.
\end{aligned}$$

For $u \in V - [n]$ as above, it holds by the Markov property, induction, and Lemma 3 that

$$\begin{aligned}
\Gamma_u^i(\zeta) &= \ln \mathbf{E}\left[\exp\left(\zeta \sum_{\alpha=1,2} S_{v_\alpha} \frac{\psi(e_\alpha)}{\theta_{e_\alpha}}\right) \mid \sigma_u = \nu_i\right] \\
&= \sum_{\alpha=1,2} \ln \mathbf{E}\left[\exp\left(\zeta S_{v_\alpha} \frac{\psi(e_\alpha)}{\theta_{e_\alpha}}\right) \mid \sigma_u = \nu_i\right] \\
&= \sum_{\alpha=1,2} \ln\left(\sum_{j\in\Phi} M_{ij}^{e_\alpha} \mathbf{E}\left[\exp\left(\zeta S_{v_\alpha} \frac{\psi(e_\alpha)}{\theta_{e_\alpha}}\right) \mid \sigma_{v_\alpha} = \nu_j\right]\right) \\
&= \sum_{\alpha=1,2} \ln\left(\sum_{j\in\Phi} M_{ij}^{e_\alpha} \exp\left(\Gamma_{v_\alpha}^j\left(\zeta \frac{\psi(e_\alpha)}{\theta_{e_\alpha}}\right)\right)\right) \\
&\leq \sum_{\alpha=1,2} \ln\left(\sum_{j\in\Phi} M_{ij}^{e_\alpha} \exp\left(\nu_j\left(\zeta \frac{\psi(e_\alpha)}{\theta_{e_\alpha}}\right) + \frac{1}{2} c K_{v_\alpha,w}\left(\zeta \frac{\psi(e_\alpha)}{\theta_{e_\alpha}}\right)^2\right)\right) \\
&= \frac{1}{2} c\zeta^2 \sum_{\alpha=1,2} K_{v_\alpha,w}\left(\frac{\psi(e_\alpha)}{\theta_{e_\alpha}}\right)^2 \\
&\qquad + \sum_{\alpha=1,2} \ln\left(\sum_{j\in\Phi} M_{ij}^{e_\alpha} \exp\left(\nu_j\left(\zeta \frac{\psi(e_\alpha)}{\theta_{e_\alpha}}\right)\right)\right) \\
&\leq \frac{1}{2} c\zeta^2 \sum_{\alpha=1,2} K_{v_\alpha,w}\left(\frac{\psi(e_\alpha)}{\theta_{e_\alpha}}\right)^2 \\
&\qquad + \sum_{\alpha=1,2} \theta_{e_\alpha}\nu_i\left(\zeta \frac{\psi(e_\alpha)}{\theta_{e_\alpha}}\right) + \frac{1}{2} c(1-\theta_{v_\alpha}^2)\left(\zeta \frac{\psi(e_\alpha)}{\theta_{e_\alpha}}\right)^2 \\
&= \nu_i\zeta + \frac{1}{2} c\zeta^2 \sum_{\alpha=1,2} ((1-\theta_{v_\alpha}^2) + K_{v_\alpha,w})\left(\frac{\psi(e_\alpha)}{\theta_{e_\alpha}}\right)^2 \\
&= \nu_i\zeta + \frac{1}{2} c\zeta^2 K_{u,w}.
\end{aligned}$$

■

Proof of Lemma 2: The expectation formula is proved first. Note that

$$\begin{aligned}
&\mathbf{E}\left[\sum_{a'\in A}\sum_{b'\in B} w(a')w(b')\Theta(a,a')^{-1}\Theta(b,b')^{-1}\widehat{\Theta}(a',b')\right] \\
&= \mathbf{E}\left[\frac{1}{k}\sum_{i=1}^{k}\left(\sum_{a'\in A}\frac{2^{-h'}\sigma_{a'}^{i}}{\Theta(a,a')}\right)\left(\sum_{b'\in B}\frac{2^{-h'}\sigma_{b'}^{i}}{\Theta(b,b')}\right)\right] \\
&= \mathbf{E}\left[\left(\sum_{a'\in A}\frac{2^{-h'}\sigma_{a'}}{\Theta(a,a')}\right)\left(\sum_{b'\in B}\frac{2^{-h'}\sigma_{b'}}{\Theta(b,b')}\right)\right] \\
&= \mathbf{E}\left[\mathbf{E}\left[\left(\sum_{a'\in A}\frac{2^{-h'}\sigma_{a'}}{\Theta(a,a')}\right)\left(\sum_{b'\in B}\frac{2^{-h'}\sigma_{b'}}{\Theta(b,b')}\right)\mid\sigma_a,\sigma_b\right]\right] \\
&= \mathbf{E}\left[\mathbf{E}\left[\sum_{a'\in A}\frac{2^{-h'}\sigma_{a'}}{\Theta(a,a')}\mid\sigma_a\right]\mathbf{E}\left[\sum_{b'\in B}\frac{2^{-h'}\sigma_{b'}}{\Theta(b,b')}\mid\sigma_b\right]\right] \\
&= \mathbf{E}\left[\sigma_a\sigma_b\right] \\
&= e^{-\delta(a,b)},
\end{aligned}$$

where it was used that $|A| = |B| = 2^{h'}$.

To prove the large deviation result, it suffices by standard arguments (*S11*) to bound the exponential moment of

$$S_aS_b = \left(\sum_{a'\in A}\frac{2^{-h'}\sigma_{a'}^{i}}{\Theta(a,a')}\right)\left(\sum_{b'\in B}\frac{2^{-h'}\sigma_{b'}^{i}}{\Theta(b,b')}\right).$$

Let N be $\mathrm{Normal}(0,1)$ and recall that $\mathbf{E}[e^{\zeta N}] = e^{\zeta^2/2}$. By applying Proposition 2 twice and using Fubini's Theorem for positive random variables, it follows that (letting w be the homoge-

neous flow on T)

$$\begin{aligned}
\mathbf{E}[\exp(\zeta S_a S_b) \,|\, \sigma_a, \sigma_b] &\leq \mathbf{E}[\exp(\sigma_a \zeta S_b + \frac{1}{2} c \zeta^2 S_b^2 K_{a,w}) \,|\, \sigma_a, \sigma_b] \\
&= \mathbf{E}[\exp(\sigma_a \zeta S_b + \sqrt{c K_{a,w}} \zeta S_b N) \,|\, \sigma_a, \sigma_b] \\
&= \mathbf{E}[\exp(S_b(\sigma_a \zeta + \sqrt{c K_{a,w}} \zeta N)) \,|\, \sigma_a, \sigma_b] \\
&\leq \mathbf{E}[\exp(\sigma_b(\sigma_a \zeta + \sqrt{c K_{a,w}} \zeta N) \\
&\qquad + \frac{1}{2} c (\sigma_a \zeta + \sqrt{c K_{a,w}} \zeta N)^2 K_{b,w}) \,|\, \sigma_a, \sigma_b] \\
&< +\infty,
\end{aligned}$$

uniformly in σ_a, σ_b for $|\zeta| > 0$ small enough, where it was used that $|\sigma_a|, |\sigma_b| \leq \bar{\nu} < +\infty$ and

$$\mathbf{E}[e^{c^2 \zeta^2 K_{a,w} K_{b,w} N^2}] = \left(\frac{1}{1 - 2(c^2 \zeta^2 K_{a,w} K_{b,w})} \right)^{1/2} < +\infty,$$

for small enough ζ. Cauchy-Schwarz and the moment-generating function of the chi-square distribution were also used.

A more careful analysis gives the dependence of χ in ε as $\varepsilon \to 0$. By standard concentration results, the value of χ is given by

$$\chi_{\sigma_a, \sigma_b} = \inf_{0 < \zeta < 1} \exp(-\zeta(\mathbf{E}[S_a S_b \,|\, \sigma_a, \sigma_b] + \varepsilon)) \mathbf{E}[\exp(\zeta S_a S_b) \,|\, \sigma_a, \sigma_b].$$

Note that

$$\mathbf{E}[S_a S_b \,|\, \sigma_a, \sigma_b] = \sigma_a \sigma_b,$$

so that the first term cancels out in the bound on $\mathbf{E}[\exp(\zeta S_a S_b) \,|\, \sigma_a, \sigma_b]$. Take $\zeta = \gamma \varepsilon$ for a

small $\gamma > 0$. By Cauchy-Schwarz,

$$\begin{aligned}
\chi_{\sigma_a,\sigma_b} &\leq \exp(-\gamma\varepsilon^2 + \frac{1}{2}c\bar{\nu}^2 K_{b,w}\gamma^2\varepsilon^2) \\
&\quad \times \Big(\mathbf{E}[\exp(2(\sigma_b + c\sigma_a K_{b,w}\gamma\varepsilon)\sqrt{cK_{a,w}}\gamma\varepsilon N) \mid \sigma_a, \sigma_b]\Big)^{1/2} \\
&\quad \times \big(\mathbf{E}[\exp(c^2 K_{a,w}K_{b,w}\gamma^2\varepsilon^2 N^2) \mid \sigma_a, \sigma_b]\big)^{1/2} \\
&\leq \exp(-\gamma\varepsilon^2 + \frac{1}{2}c\bar{\nu}^2 K_{b,w}\gamma^2\varepsilon^2) \\
&\quad \times \big(\exp(2(\sigma_b + c\sigma_a K_{b,w}\gamma\varepsilon)^2 cK_{a,w}\gamma^2\varepsilon^2 N)\big)^{1/2} \\
&\quad \times \left(\frac{1}{1 - 2(c^2\gamma^2\varepsilon^2 K_{a,w}K_{b,w})}\right)^{1/4} \\
&= 1 - (\gamma - \frac{1}{2}c\bar{\nu}^2 K_{b,w}\gamma^2 - c\bar{\nu}^2 K_{a,w}\gamma^2 - \frac{1}{2}c^2 K_{a,w}K_{b,w}\gamma^2)\varepsilon^2 + O(\varepsilon^3).
\end{aligned}$$

Taking γ small enough as a function of c, $\bar{\nu}$, $K_{a,w}$, and $K_{b,w}$, one can make the parenthesis above positive and one finally gets

$$\chi_{\sigma_a,\sigma_b} = 1 - \Omega(\varepsilon^2),$$

for all σ_a, σ_b.

To prove that the large deviation result is independent of the level h', it is shown that $K_{a,w}$ is uniformly bounded in h'. From (10),

$$\begin{aligned}
K_{a,w} &\leq \sum_{i=0}^{h'-1}(1 - e^{-2g})2^{h'-i}\frac{e^{2(h'-i)g}}{2^{2(h'-i)}} \\
&\leq \sum_{j=1}^{h'} e^{2jg}e^{-(2\ln\sqrt{2})j} \\
&= \sum_{j=1}^{h'} e^{2j(g-g^{**})} \\
&\leq \sum_{j=0}^{+\infty}(e^{-2(g^{**}-g)})^j \\
&= \frac{1}{1 - e^{-2(g^{**}-g)}} < +\infty, \qquad (14)
\end{aligned}$$

where recall that $g^{**} = \ln \sqrt{2}$ and $g < g^{**}$. ■

Proof of Proposition 3: Let

$$\varepsilon' = \min\{(e^{\varepsilon} - 1)e^{-D}, (1 - e^{-\varepsilon})e^{-D}\},$$

and observe that

$$\begin{aligned}
&\bar{\delta}_{\mathrm{c}}(a, b) - \delta(a, b) < -\varepsilon \\
&\implies e^{-\bar{\delta}_{\mathrm{c}}(a,b)} > e^{-\delta(a,b)+\varepsilon} \\
&\implies e^{-\bar{\delta}_{\mathrm{c}}(a,b)} - e^{-\delta(a,b)} > (e^{\varepsilon} - 1)e^{-D} \geq \varepsilon'.
\end{aligned}$$

A similar implication holds in the other direction. The result now follows from Lemma 2. Note that the dependence of κ in ε comes from the fact that $\chi = 1 - \Omega((\varepsilon')^2)$ and $\varepsilon' = O(\varepsilon)$ as $\varepsilon \to 0$. ■

Proof of Proposition 4: The proof is similar to the proof of Proposition 3. ■

Proof of Proposition 5: Let $\varepsilon < \Delta/2$. The first part of the proposition follows immediately from Proposition 3 and the second part of Proposition 4. For the second part, choose κ so as to satisfy the conditions of Proposition 3 with diameter $D + \ln W$ and apply the first part of Proposition 4 using the remarks above the statement of Proposition 5. The dependence of κ in Δ follows immediately from the corresponding statement in regarding ε in Proposition 3. ■

Supplemental References

S1. K. Tamura, M. Nei, *Mol Biol Evol* **10**, 512 (1993).

S2. J. Felsenstein, *Inferring Phylogenies* (Sinauer, Sunderland, MA, 2004).

S3. A. Rzhetsky, T. Sitnikova, *Mol Biol Evol* **13**, 1255 (1996).

S4. E. Mossel, Y. Peres, *Ann. Appl. Probab.* **13**, 817 (2003).

S5. W. S. Evans, C. Kenyon, Y. Peres, L. J. Schulman, *Ann. Appl. Probab.* **10**, 410 (2000).

S6. S. Roch, *FOCS'08: Annual IEEE Symposium on Foundations of Computer Science* (IEEE Computer Society, Los Alamitos, CA, USA, 2008), pp. 729–738.

S7. C. Daskalakis, E. Mossel, S. Roch, Evolutionary trees and the Ising model on the Bethe lattice: a proof of Steel's conjecture (2009). *Probab Theor Relat Field (In Press).*

S8. X. Gu, W. H. Li, *Proceedings of the National Academy of Sciences of the United States of America* **93**, 4671 (1996).

S9. X. Gu, W.-H. Li, *Proceedings of the National Academy of Sciences of the United States of America* **95**, 5899 (1998).

S10. L. Wasserman, *All of statistics*, Springer Texts in Statistics (Springer-Verlag, New York, 2004). A concise course in statistical inference.

S11. R. Durrett, *Probability: theory and examples* (Duxbury Press, Belmont, CA, 1996), second edn.

Algorithm WPGMA
Input: Distance estimates $\{\hat{\delta}_{\mathrm{u}}(a,b)\}_{a,b\in[n]}$;
Output: Tree;

- **Initialization.** Let $\mathcal{Z}_0$ be the set of leaves as clusters, that is,

$$\mathcal{Z}_0 = \{\{l\} \ : \ l \in [n]\},$$

and for all $a, b \in [n]$ let

$$\bar{\delta}_{\mathrm{u}}(\{a\},\{b\}) = \hat{\delta}_{\mathrm{u}}(a,b).$$

- **Main Loop.** For $i = 1, \ldots, n-1$,
  - **Selection Step.** Let

$$(A^*, B^*) \in \arg\min\{\bar{\delta}_{\mathrm{u}}(A,B) \ : \ A, B \in \mathcal{Z}_{i-1} \text{ distinct}\}.$$

Merge clusters A^*, B^* to obtain $\mathcal{Z}_i$.

  - **Reduction Step.** For all $C \in \mathcal{Z}_i - \{A^* \cup B^*\}$, compute

$$\bar{\delta}_{\mathrm{u}}(C, A^* \cup B^*) = \frac{1}{2}[\bar{\delta}_{\mathrm{u}}(C,A^*) + \bar{\delta}_{\mathrm{u}}(C,B^*)]. \tag{15}$$

- **Output.** Output tree implied by the successive clusterings $\mathcal{Z}_0, \ldots, \mathcal{Z}_{n-1}$.

Figure 1: Algorithm WPGMA.

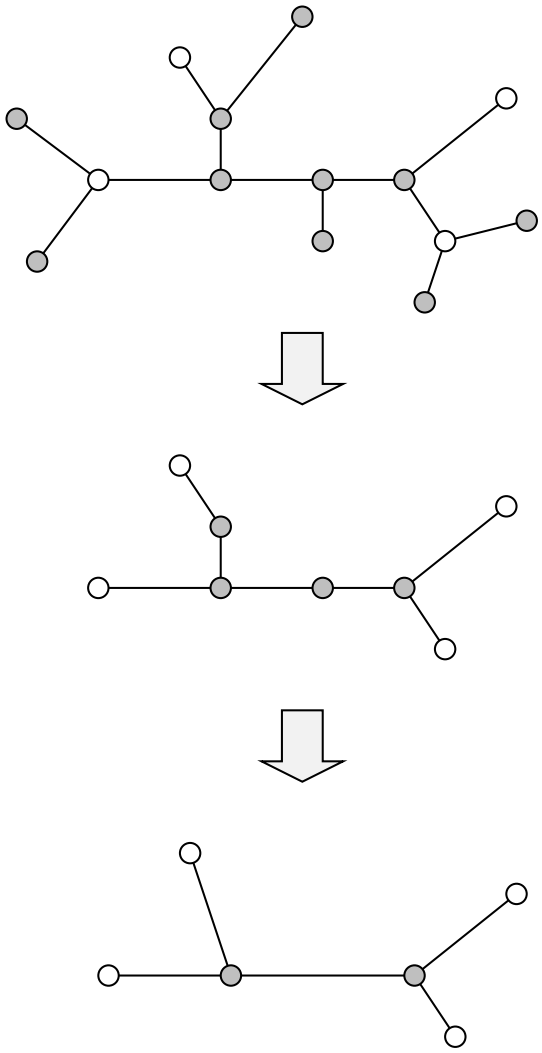

Figure 2: Restricting the top tree to its white nodes.

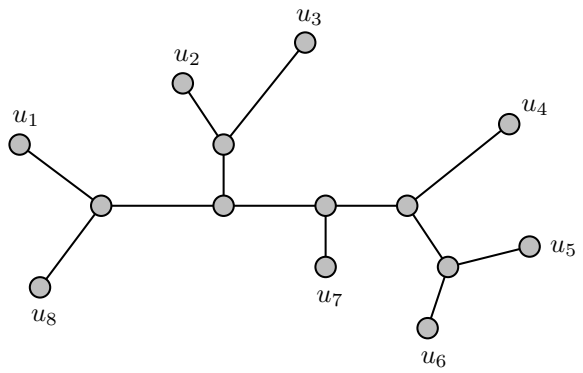

Figure 3: The subtrees $T|_{\{u_1,u_2,u_3,u_8\}}$ and $T|_{\{u_4,u_5,u_6,u_7\}}$ are edge-disjoint. The subtrees $T|_{\{u_1,u_5,u_6,u_8\}}$ and $T|_{\{u_2,u_3,u_4,u_7\}}$ are edge-sharing.

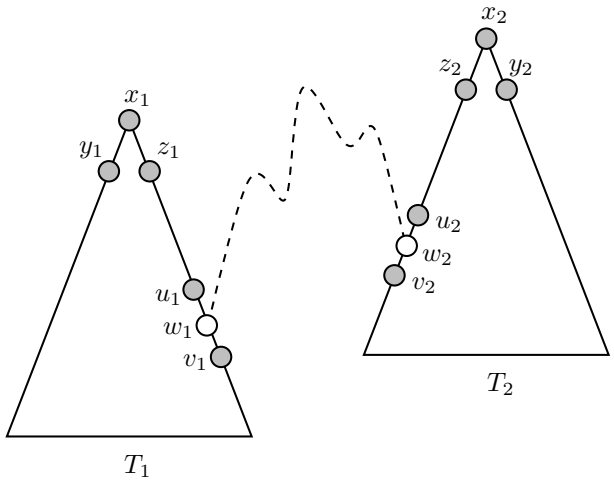

Figure 4: Basic Disjoint Setup (General). The rooted subtrees T_1, T_2 are edge-disjoint but are not assumed to be dangling. The white nodes may not be in the restricted subtrees T_1, T_2. The case $w_1 = x_1$ and/or $w_2 = x_2$ is possible. Note that if one roots the tree at any node along the dashed path, the subtrees rooted at y_1 and y_2 are edge-disjoint and dangling (unlike T_1 and T_2).

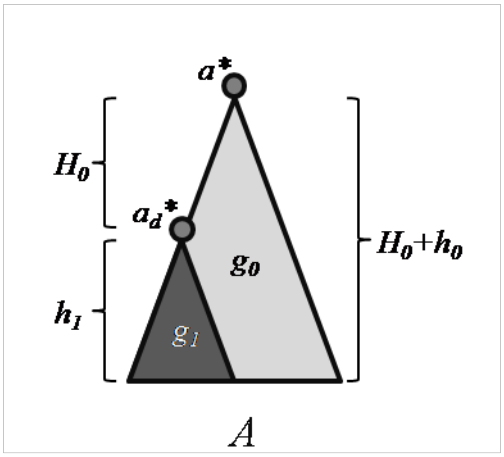

Figure 5: Illustrative example (not to scale). The g_0 and g_1 values are branch lengths. The H_0, h_0, and h_1 values indicate numbers of levels.

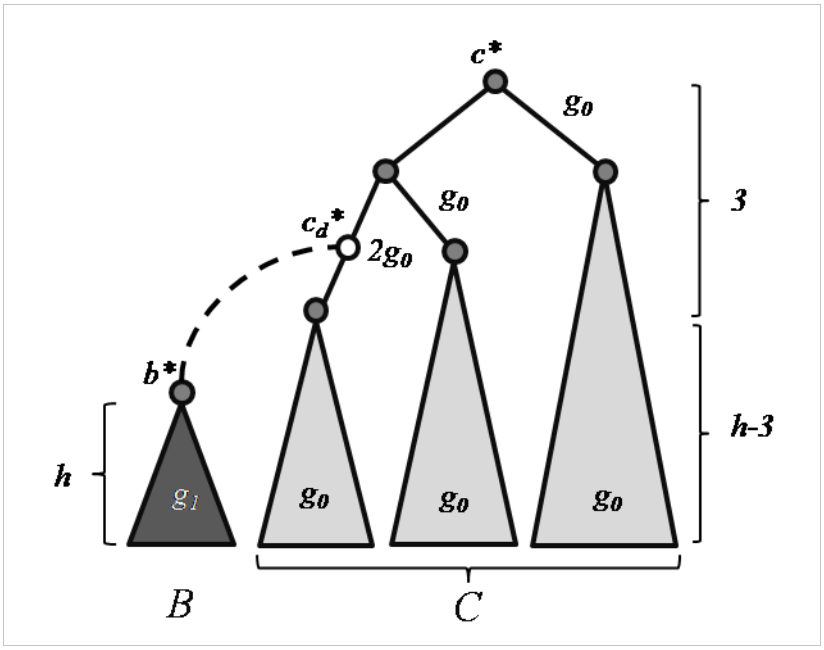

Figure 6: In the terminology of the Basic Disjoint Setup (General), $x_1 = b^*$ and y_1, z_1 are the direct descendants of b^* in B. Similarly, $x_2 = c^*$ and y_2, z_2 are the direct descendants of c^* in C. The node w_1 coincides with x_1 (and in that case u_1 and v_1 are not relevant). The node w_2 coincides with c^*_d, which is also the parent of a^*_d in A.

Algorithm DISTORTEDMETRIC
Input: Rooted forest $\mathcal{F} = \{T_1, T_2\}$ rooted at vertices x_1, x_2; weights $\delta(e)$, for all $e \in \mathcal{E}(T_1) \cup \mathcal{E}(T_2)$;
Output: Distance Υ;

- [Children] Let y_ι, z_ι be the children of x_ι in $\mathcal{F}$ for $\iota = 1, 2$ (if x_ι is a leaf, set $z_\iota = y_\iota = x_\iota$);
- [Distance Computations] For all pairs $(a, b) \in \{y_1, z_1\} \times \{y_2, z_2\}$, compute

$$\mathcal{D}(a, b) := \bar{\mathrm{d}}_{\mathrm{c}}(a, b) - \delta(a, x_1) - \delta(b, x_2);$$

- [Multiple Test] If

$$\max\{\Big|\mathcal{D}(r_1^{(1)}, r_2^{(1)}) - \mathcal{D}(r_1^{(2)}, r_2^{(2)})\Big| : (r_1^{(\iota)}, r_2^{(\iota)}) \in \{y_1, z_1\} \times \{y_2, z_2\}, \iota = 1, 2\} = 0,$$

return $\Upsilon := \mathcal{D}(z_1, z_2)$, otherwise return $\Upsilon := +\infty$ (return $\Upsilon := +\infty$ if any of the distances above is $+\infty$).

Figure 7: Routine DISTORTEDMETRIC.

Algorithm BLINDFOLDED CHERRY PICKING (distance-based version)
Input: Similarity estimates $\{\widehat{\Theta}((,a),b)\}_{a,b\in[n]}$;
Output: Estimated topology;

- **0) Initialization:**
  - [Iteration Counter] $i := 0$;
  - [Rooted Subforest] $\mathcal{F}_0 := [n]$;
  - [Local Metric] For all $u, v \in \mathcal{F}_0$, set $\mathcal{D}_0(u,v) = \bar{\mathrm{d}}_{\mathrm{c}}(u,v)$;
- **1) Local Cherry Identification:**
  - Iteration: i;
  - Set $\mathcal{F}_{i+1} := \mathcal{F}_i$;
  - For all $(v_1, w_1) \in \binom{\rho(\mathcal{F}_i)}{2}$,
    - [Main Step] $(\mathrm{IsCherry}, l_v, l_w) :=$ LOCALCHERRY $((v_1,w_1);(\mathcal{F}_i,\mathcal{D}_i))$;
    - If IsCherry = TRUE,
      - [Update Forest] Create new node u_1 and add cherry (v_1,u_1,w_1) to $\mathcal{F}_{i+1}$;
      - [Edge Lengths] Set $h(u_1,v_1) := l_v$ and $h(u_1,w_1) := l_w$;
- **2) Collision Removal:**
  - [Update Metric] For all $x_1, x_2 \in \rho(\mathcal{F}_{i+1})$, for all $u_\iota \in T_{x_\iota}^{\mathcal{F}_{i+1}}$, $\iota = 1,2$, set
    $$\mathcal{D}_{i+1}(u_1,u_2) = \text{DISTORTEDMETRIC}(u_1,u_2;\mathcal{F}_{i+1};\{h(e)\}_{e\in\mathcal{F}_{i+1}});$$
  - Set $\mathcal{F} := \mathcal{F}_{i+1}$ and $\mathcal{D} := \mathcal{D}_{i+1}$;
  - For all $(u_0,u_1) \in \rho(\mathcal{F}) \times \rho(\mathcal{F})$,
    - Set HasCollision := FALSE;
    - If u_1 is not a leaf,
      - [Main Step] $(\mathrm{HasCollision}, z) :=$ DETECTCOLLISION $((u_0,u_1);(\mathcal{F},\mathcal{D}))$;
    - If HasCollision = TRUE,
      - [Update Forest] $\mathcal{F}_{i+1} :=$ REMOVECOLLISION$(z;\mathcal{F}_{i+1})$;
  - [Second Pass] Set $\mathcal{F} := \mathcal{F}_{i+1}$ and repeat the previous step;
- **3) Termination:**
  - If $|\rho(\mathcal{F}_{i+1})| \le 3$,
    - Join nodes in $\rho(\mathcal{F}_{i+1})$ (star if 3, single edge if 2);
    - Return (tree) $\mathcal{F}_{i+1}$;
  - Else, set $i := i+1$, and go to Step 1.

Figure 8: Algorithm BLINDFOLDED CHERRY PICKING.

Algorithm LOCALCHERRY
Input: Two nodes (v_1, w_1); current forest and distance matrix $(\mathcal{F}, \mathcal{D})$;
Output: Boolean value and length estimates;

- Set IsCherry := TRUE and $l_v = l_w = 0$;
- [Short Distance]
  - If $\mathcal{D}(v_1, w_1) > 2g$, then IsCherry := FALSE;
- [Local Cherry]
  - Set $\mathcal{N} = \left\{(v_2, w_2) \in \binom{\rho(\mathcal{F})}{2} : \overline{\mathcal{D}}(\{v_1, w_1, v_2, w_2\}) \leq 5g\right\}$;
  - If $\mathcal{N}$ is empty, then IsCherry := FALSE; Else, for all $(v_2, w_2) \in \mathcal{N}$,
    - If ISSPLIT $((v_1, w_1), (v_2, w_2); \mathcal{D}) =$ FALSE then set IsCherry := FALSE and *break*;
- [Edge Lengths]
  - If IsCherry = TRUE,
    - Let x_1, x_2 be the children of v_1 in $\mathcal{F}$ (or let $x_1 = x_2 = v_1$ if v_1 is a leaf);
    - Let z_0 be the closest node to v_1 in $\rho(\mathcal{F}) - \{v_1, w_1\}$ under $\mathcal{D}$;
    - Set $(b_v, l_v) :=$ ISSHORT $((x_1, x_2), (w_1, z_0); \mathcal{F})$;
    - Repeat previous steps switching the roles of v_1 and w_1;
    - Set IsCherry := $b_v \wedge b_w$;
- Return $(\text{IsCherry}, l_v, l_w)$;

Figure 9: Routine LOCALCHERRY.

Algorithm DETECTCOLLISION
Input: Two roots u_0, u_1; directed forest and distance matrix $(\mathcal{F}, \mathcal{D})$;
Output: Boolean value and node;

- Set $\mathrm{HasCollision} :=$ FALSE and $z := 0$;
- Let x_0, y_0 be the children of u_0 in $\mathcal{F}$;
- Scan through all nodes v in T_{u_1} (except u_1) in reverse BFS manner,
  - Let $w := \mathrm{Sister}_{\mathcal{F}}(v)$ and $u := \mathrm{Parent}_{\mathcal{F}}(v)$;
  - [Collision Test] Compute

$$b_x := \textsc{IsCollision}(x_0, v, w, u; h(u, v); (\mathcal{F}, \mathcal{D})),$$

    and

$$b_y := \textsc{IsCollision}(y_0, v, w, u; h(u, v); (\mathcal{F}, \mathcal{D}));$$

  - If $\mathrm{HasCollision} := b_x \wedge b_y =$ TRUE then set $z := v$ and *break*;
- Return $(\mathrm{HasCollision}, z)$;

Figure 10: Routine DETECTCOLLISION.

Algorithm REMOVECOLLISION
Input: Node v; rooted forest $\mathcal{F}$;
Output: Rooted forest;

- If v is not in $\mathcal{F}$ or v is a root in $\mathcal{F}$, return $\mathcal{F}$;
- Let z_0 be the root of the subtree of $\mathcal{F}$ in which v lies;
- Set $x := v$;
- While $x \neq z_0$,
  - Set $x := \mathrm{Parent}_{\mathcal{F}}(x)$;
  - Remove node x and its adjacent edges below it;
- Return the updated forest $\mathcal{F}$.

Figure 11: Routine REMOVECOLLISION.

Algorithm ISSHORT
Input: Two pairs of nodes (v_1, w_1), (v_2, w_2); rooted forest $\mathcal{F}$;
Output: Boolean value and length estimate;

- [Internal Edge Length] Set

$$\nu = \frac{1}{2}(\bar{\mathrm{d}}_{\mathrm{c}}(v_1, v_2) + \bar{\mathrm{d}}_{\mathrm{c}}(w_1, w_2) - \bar{\mathrm{d}}_{\mathrm{c}}(v_1, w_1) - \bar{\mathrm{d}}_{\mathrm{c}}(v_2, w_2));$$

- [Test] If $\nu \leq g$ return $(\textsc{true}, \nu)$, o.w. return $(\textsc{false}, 0)$;

Figure 12: Routine ISSHORT.

Algorithm IsSplit
Input: Two pairs of nodes (v_1, w_1), (v_2, w_2); a distance matrix $\mathcal{D}$ on these four nodes;
Output: TRUE or FALSE;

- [Internal Edge Length] Set

$$\nu = \frac{1}{2}\left(\mathcal{D}(w_1, w_2) + \mathcal{D}(v_1, v_2) - \mathcal{D}(w_1, v_1) - \mathcal{D}(w_2, v_2)\right);$$

  (set $\nu = +\infty$ if any of the distances is $+\infty$)

- [Test] If $\nu < f/2$ return FALSE, o.w. return TRUE.

Figure 13: Routine IsSplit.

Algorithm IsCollision
Input: Four nodes x_0, v, w, u; an edge length h; a rooted forest and a distance matrix $(\mathcal{F}, \mathcal{D})$;
Output: TRUE or FALSE;

- [Children] Let v_1, v_2 be the children of v in $\mathcal{F}$ (or $v_1 = v_2 = v$ if v is a leaf);
- [Internal Edge Length] Set

$$\nu = \frac{1}{2}\left(\mathcal{D}(v_1, x_0) + \mathcal{D}(v_2, w) - \mathcal{D}(v_1, v_2) - \mathcal{D}(x_0, w)\right);$$

  (set $\nu = +\infty$ if any of the distances is $+\infty$)

- [Test] If $(h - \nu) > f/2$ return TRUE, else return FALSE;

Figure 14: Routine IsCollision.