Inferring ancestral sequences in taxon-rich phylogenies

Olivier Gascuel, Mike Steel

Inferring ancestral sequences in taxon-rich phylogenies

Olivier Gascuel[1], Mike Steel[2]

[1] *Méthodes et Algorithmes pour la Bioinformatique, LIRMM, CNRS - Université de Montpellier, 161 rue Ada, 34392 Montpellier, FRANCE Tel: +33 (0) 4 67 41 85 47; Fax: +33 (0) 4 67 41 85 00; Email: gascuel@lirmm.fr*

[2] *Corresponding author*
Allan Wilson Centre for Molecular Ecology and Evolution University of Canterbury, Christchurch, NEW ZEALAND Tel: +64 3 3667001; Fax: +64-3-3642587; Email: m.steel@math.canterbury.ac.nz

Abstract

Statistical consistency in phylogenetics has traditionally referred to the accuracy of estimating phylogenetic parameters for a fixed number of species as we increase the number of characters. However, it is also useful to consider a dual type of statistical consistency where we increase the number of species, rather than characters. This raises some basic questions: what can we learn about the evolutionary process as we increase the number of species? In particular, does having more species allow us to infer the ancestral state of characters accurately? This question is particularly important when sequence evolution varies in a complex way from character to character, as methods applicable for i.i.d. models may no longer be valid. In this paper, we assemble a collection of results to analyse various approaches for inferring ancestral information with increasing accuracy as the number of taxa increases.

Keywords: ancestral sequence, Markov process, parsimony

1. Introduction

As Elliott Sober discussed two decades ago [15], there is a fundamental asymmetry between reconstructing a past state from a present observation, and predicting its future state. Moreover, this holds even when the state evolves according to a time-reversible process (processes which, when they are in equilibrium, behave the same whether run forward or backward in time). For instance, consider any continuous Markov process on two states, with arbitrary transition rates (generally unequal) between the two states. If we observe the state of the process at the present time t, then the 'best' estimate of the initial state at time 0 is always the present state, but the 'best' estimate of its state at some future time $t' > t$ depends on the actual transition rates (which may be unknown) [15].

When we move beyond two states in a Markov process, the current state is no longer guaranteed to always be the 'best' estimate of the ancestral state, even for reversible processes, as we describe below. Ancestral state estimation assumes a further dimension when we move from the linear evolution of a state through time to the bifurcating evolution of states in a tree that results in their observed values at the leaves. The presence of many leaves helps us to estimate the ancestral state more accurately, but these leaves do not provide independent information about the root state due to correlations arising from the partial overlap of the paths in the tree as one moves from the root to the leaves. The mathematical, statistical and computational aspects of ancestral state estimation on a tree have been explored by a number of authors (e.g. [5, 8, 11, 12, 13, 14, 22]) and the inference of ancestral states is an important question in biology [9].

Our interest here is in site-specific models. These are especially relevant with proteins, where each site has specific biochemical constraints (e.g. small and hydrophobic, aromatic, helix-former, etc). As we are interested in site-specific models, the details of the substitution model are mostly unknown. For example, the relative or absolute branch length may not be known exactly, though we may have some upper bound on them. Also, the equilibrium frequencies at the site may not be known. This is the case in the CAT model for proteins ([7]; see also [6]). This model is a mixture of F81-like models, where each site follows a Poisson model with specific amino acid frequencies defined by the biochemical constraints acting on that site. However, we

shall see that dealing with unknown equilibrium frequencies imposes strong limitations when the aim is to estimate ancestral character states, especially when the branch lengths are unknown. Thus, we will also envisage special cases where equilibrium frequencies are known or even all identical.

In most cases (e.g. when the branch lengths are unknown), we are thus unable to use standard likelihood calculations based on the pruning algorithm to compute the most likely character state at the tree root. Thus, we will discuss and study simple decision rules to predict the state at the tree root. Parsimony is an example of such a rule, where the branch lengths are not used. Another example is the majority rule that involves selecting the state that is most frequent at the tree leaves to estimate the root state. For models in which the equilibrium frequencies are not uniform across states, more complex inference rules are required. We shall see that under suitable assumptions on the tree topology and branch lengths and/or on the model, these simple rules are statistically consistent as we increase the number of taxa.

We treat four general cases, each depending on the properties of the model. We start with the simplest model, the *symmetric Poisson model* in which all transition rates between distinct rates are equal (in the case of four states, this is the well-known Jukes-Cantor model). We then consider two overlapping generalizations ('monotone' and 'conservative') and finally we deal with the general model, for which stronger assumptions on the tree are required.

1.1. Preliminaries

Consider a rooted phylogenetic tree T (possibly non-binary) with n leaves and a set $\mathcal{S}$ of possible states that each vertex can be in. For a single-site assignment of states at the leaves of T, assume that the assignment has evolved under a GTR (general time-reversible) model from a particular character state s_0 at the root, with a normalized rate matrix $Q = \Pi S$ (where $\Pi = \text{diag}(\pi)$ contains the equilibrium frequencies, and S is a symmetric matrix of 'exchangeabilities'). The process acts on each edge e according to some associated branch length l_e.

We assume that T and S (and perhaps π) are given, and, in addition,

we may either know l_e or have some bounds on them (e.g. the sum of the lengths from the root to any tip is, at most, some given value l). We would like to use this input to estimate the ancestral state $s_0 \in \mathcal{S}$ at the root of the tree.

The ability to estimate s_0 accurately depends on a tradeoff between what we know about the underlying parameters (e.g. the site rate parameter μ, the branch lengths l_e, and the properties of Q such as the equilibrium distribution π) and how 'well behaved' the underlying Markov process is.

In particular, we seek a method M that is statistically consistent in the following sense: Suppose that the character states at the leaves have evolved from an unknown state at some ancestral root vertex under some Markov model. Then M is *statistically consistent* as an estimator of the root state, from character state data at the leaves of the tree, if the probability that M returns the correct ancestral state is at least $1 - g(n, \xi)$ where g is some function which tends to zero as $\max\{\frac{1}{n}, \xi\}$ tends to zero, n is the number of leaves of the tree, and ξ is a parameter describing constraints on the branch lengths of the tree.

A natural choice of such a method, when Q is completely specified (including the equilibrium distribution π) and the branch lengths (l_e) are also known exactly, is to take the maximum posterior probability (MPP) ancestral state (this selects the state with the largest posterior probability; the MPP method can be shown to confer the largest expected correct reconstruction probability amongst all methods). For a symmetric model with flat priors the MPP estimate of the root state is the same as the maximum likelihood (ML) estimate, but in general the two approaches differ.

When the model is (partly) unknown, the ML and MPP approaches may not be possible (since they require both the tree topology and estimates of the model parameters). But in these cases, simpler approaches exist. For example, for a simple symmetric model (e.g. Jukes-Cantor) and a star tree with unknown branch lengths that are bounded above ($l_e \leq l < \infty$), we can estimate the ancestral state accurately by selecting the majority state (the consistency of this approach is justified by large deviation theorems for sums of independent random variables).

4

However, even for symmetric models, it is clear that simply allowing n to grow is not sufficient to allow for accurate inference of the ancestral state s_0; for example, we could have just two long edges incident with the root, and lots of very short edges that join the other endpoints of these edges to numerous taxa. In this case, the substitution process behaves almost as on a two-taxon tree and we have little information on the root when the two branches become too long. Thus we seek relevant and reasonable constraints on the distribution of l_e values for this accurate estimation to be possible.

Moving away from symmetric models, selecting the majority state at the leaves as an estimate of the ancestral state is not generally a sound strategy, even for a star tree, since the process after a long period of time will favour the state with the highest equilibrium frequency, regardless of the state at the root.

Although we deal with the inference of a state at a single site, the results are still relevant to the more general question of ancestral reconstruction of a sequence (of length k) from sequences of length k observed at the leaves of the tree. Assuming independent site evolution, the problem of ancestral state estimation remains the same (i.e. each site is solved independently). If, on the other hand, sites evolve with dependencies, but subject to some Markov process, then the sequences of length k (small) may be treated as single character states in a larger state space.

2. Case I: Root state estimation without detailed knowledge of l_e under a symmetric Poisson model

Under the symmetric r-state Poisson model, the maximum likelihood estimate of the root state, in the case where the branch lengths (l_e) are unknown and are regarded as nuisance parameters to be optimized, is the maximum parsimony (MP) estimate (Theorem 6 of [21]). In this setting, we can reliably estimate the root state, provided the taxon sampling is sufficiently dense that no edges are too long. This was suggested by the simulations in [14] and we establish two formal results now for the case when $r = 2$.

Proposition 2.1. *Consider any rooted binary phylogenetic tree T. Evolve a single site under the two-state symmetric Poisson model. Let l_+ be the*

maximum branch length over all edges. Provided that $l_+ < \frac{1}{2}\log(\frac{4}{3})$, the probability P^* that the maximum parsimony (MP) reconstruction of the root state is the true state (toss a fair coin if the two states are equally favored) satisfies:

$$P^* \geq 1 - 4l_+.$$

Proof. When l_+ satisfies satisfies the bound described then, for each edge e of T the probability that the endpoints of edge e are in different states $p(e) = \frac{1}{2}(1 - e^{-2l_e})$ satisfies the inequality $p(e) < \frac{1}{8}$. It then follows from part (ii) of Lemma 5.1 of [**?**], that:

$$P^* \geq \frac{1}{2} + \Delta_g,$$

where:

$$\Delta_g = \frac{\sqrt{(1 - 4g)(1 - 8g)}}{2(1 - 2g)^2},$$

and where $g = \max_e\{p(e)\}$. The result now follows from the inequalities:

$$\frac{\sqrt{(1 - 4g)(1 - 8g)}}{2(1 - 2g)^2} \geq \frac{1}{2}(1 - 8g), \text{ and } g \leq l_+.$$

$\square$

Unfortunately, in a Yule tree of fixed height, the expected value of l_+ does not converge to zero as the speciation rate λ tends to infinity. This may seem surprising, since the expected length of a randomly selected edge in the tree converges in length to 0 as λ grows; however, the expected number of edges increases with λ, and the probability that at least one of them is 'long' turns out to be positive. Simulations suggest that the expected value of l_+ converges to a value close to 60% of the height of the tree; the following result, the proof of which is provided in the Appendix, establishes a smaller lower bound.

Proposition 2.2. *Suppose a random rooted binary tree T_λ is generated by a Yule (pure birth) process with speciation rate λ acting for time t. Let $l_+ = l_+(\lambda)$ denote the length of the longest edge in T. Then $\mathbb{E}[l_+(\lambda)]$ does not converge to 0 as $\lambda \to \infty$.*

6

Thus we cannot directly apply Proposition 2.1 to Yule trees. Nevertheless, we can precisely determine the probability with which MP will correctly reconstruct the root state of a Yule tree under a symmetric Poisson substitution model on two states. In particular, provided the speciation rate passes a critical threshold (six times the substitution rate), then even for large trees where many leaves are far from the root, ancestral reconstruction becomes accurate. Moreover, as the ratio of speciation rate to substitution rate tends to infinity, we can correctly infer the root state with probability tending to 1.

2.1. MP root estimation for Yule trees under the two-state model

Consider a pure-birth Yule tree that starts with a single (root) lineage at time 0 and is grown until time t, with speciation rate λ. Suppose we also have a binary character that evolves from some ancestral state at the root of the tree towards the leaves by undergoing substitution along the edges of the tree at rate μ according to a symmetric Markov process. Thus, we have a random tree (with a random number of leaves at time t) and a random binary character observed at the leaves. Let P_t denote the probability that, on the tree produced by this Yule process at time t, the maximum parsimony estimate for the state at the root of the tree, derived from the observed states at the leaves of the tree, matches the true root state (in the case that both states are equally parsimonious at the root, select one state with equal probability).

Theorem 2.3. *Consider a random binary tree generated by a pure-birth Yule process with speciation rate λ for time t, and a binary character that evolves on this tree according to a symmetric Poisson model with substitution rate μ. Then the probability P_t that MP reconstructs on this tree the correct ancestral state from the states observed at the leaves is described as follows, where $\rho = \mu/\lambda$.*

(i) *If $\lambda > 6\mu$, then for all $t \geq 0$:*

$$\lim_{t \to \infty} P_t = \frac{1}{2}(1 + \sqrt{(1 - 6\rho)(1 - 2\rho)}) > 1 - 3\rho.$$

In particular, $\lim_{t \to \infty} P_t \to 1$ as $\rho \to 0$.

(ii) If $\lambda < 6\mu$ we have:

$$\lim_{t\to\infty} P_t = \frac{1}{2}.$$

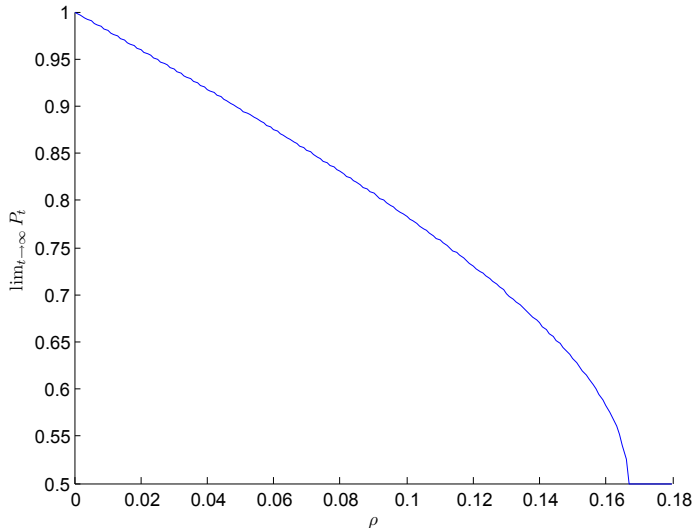Figure 1: The limiting value $\lim_{t\to\infty} P_t$ as a function of ρ for $\rho \leq 1/6$.

Proof. Let 0 and 1 denote the two states that undergo substitution on the Yule tree. Since the Markov process is symmetric we may suppose, without loss of generality, that 0 is the initial character state at time $t = 0$. From the (random) evolved states on the leaves, estimate the root state using the maximum parsimony criterion (i.e. select the root state that minimizes the total number of substitutions required to describe the evolution of the character on the tree). There may be a unique reconstructed root state (which may be the same or opposite to the true initial state) or both states may be equally parsimonious. Let S_t (resp. D_t) be the probability that 0 (resp. 1) is the unique most parsimonious root state reconstructed from the observed states at the leaves. Let $E_t = 1 - S_t - D_t$ be the probability that both states are equally parsimonious. We have:

$$P_t = S_t + \frac{1}{2}E_t = \frac{1}{2} + \frac{1}{2}(S_t - D_t). \tag{1}$$

8

We can generate a system of non-linear first-order differential equations for (S_t, D_t, E_t) as follows. Consider that in the first δ period of time, the root lineage can either:

- persist, without a substitution occurring,

- persist, with a substitution occurring, or

- it can speciate into two lineages.

This gives:

$$S_{t+\delta} = (1 - \mu\delta - \lambda\delta)S_t + \mu\delta D_t + \lambda\delta(S_t^2 + 2S_t E_t) + O(\delta^2).$$

Similarly,

$$D_{t+\delta} = (1 - \mu\delta - \lambda\delta)D_t + \mu\delta S_t + \lambda\delta(D_t^2 + 2D_t E_t) + O(\delta^2),$$

$$E_{t+\delta} = (1 - \mu\delta - \lambda\delta)E_t + \mu\delta E_t + \lambda\delta(E_t^2 + 2S_t D_t) + O(\delta^2).$$

Rearranging these expressions and letting $\delta \to 0$ produces the differential equation system:

$$\frac{dS_t}{dt} = -(\lambda + \mu)S_t + \mu D_t + \lambda(S_t^2 + 2S_t E_t);$$

$$\frac{dD_t}{dt} = -(\lambda + \mu)D_t + \mu S_t + \lambda(D_t^2 + 2D_t E_t);$$

and

$$\frac{dE_t}{dt} = -\lambda E_t + \lambda(E_t^2 + 2S_t D_t).$$

Notice that we can use the relationship $S_t + D_t + E_t = 1$ to eliminate E_t, and by writing $u = \lambda t$ we obtain a two-dimensional autonomous differential equation system for $S = S_u, D = D_u$:

$$\frac{dS}{du} = f(S, D); \quad \frac{dD}{du} = f(D, S),$$

where:

$$f(x, y) = (1 - \rho)x + \rho y - 2xy - x^2.$$

9

From the starting condition $(S, D) = (1, 0)$, at $u = t = 0$, we have:

$$S - D \geq 0 \tag{2}$$

$$S + D \geq \frac{1}{2} \tag{3}$$

These inequalities hold even when we condition on a given tree topology (for which these stronger inequalities appear as Corollary 7.21 of in [17]; see also [23], Theorem 3.1 and Corollary 3.1). Thus, (S, D) is confined to the simply-connected, two-dimensional, compact region $S, D \geq 0, \frac{1}{2} \leq S + D \leq 1, S - D \geq 0$. The equilibrium values for (S, D) are obtained by solving the system $\frac{dS}{du} = \frac{dD}{du} = 0$ which is equivalent to solving the pair of simultaneous quadratic equations $f(s, d) = 0, f(d, s) = 0$. Subtracting the second of these equations from the first gives:

$$(s - d)(1 - 2\rho - s - d) = 0. \tag{4}$$

Thus, either $s = d$ or $s + d = 1 - 2\rho$. If $s = d$, then the equation $f(s, d) = 0$ becomes $s - 3s^2 = 0$, which has a unique solution satisfying constraint (3), namely $s = d = \frac{1}{3}$.

In the other case, where $s + d = 1 - 2\rho$, $f(s, d) = 0$ becomes:

$$s^2 - (1 - 2\rho)s + \rho(1 - 2\rho) = 0,$$

which has only one solution that satisfies constraint (2), namely:

$$s = \frac{1 - 2\rho + \sqrt{(1 - 6\rho)(1 - 2\rho)}}{2}; d = \frac{1 - 2\rho - \sqrt{(1 - 6\rho)(1 - 2\rho)}}{2}. \tag{5}$$

Now, the eigenvalues of the Jacobian matrix for either equilibrium value (s, d) are $a \pm \sqrt{a^2 - b}$ where

$$a = (1 - \rho) - 2(s + d); b = a^2 - (\rho - 2s)(\rho - 2d).$$

For the first equilibrium value $(s, d) = (\frac{1}{3}, \frac{1}{3})$, both the Jacobian eigenvalues have strictly negative real parts – and so (s, d) is asymptotically stable – if and only if $\rho > \frac{1}{6}$. For the second fixed point, given by Eqn. (5), both eigenvalues have strictly negative real parts if and only if $\rho < \frac{1}{6}$. Finally, (S, D) has no limit cycle in the region described, by virtue of Dulac's criterion (with $g(x, y) = 1/xy$, for details see [19]). The results stated in the Theorem 2.3 now follow from Eqn. (1) (and, for part (i), the inequality $\frac{1}{2}(1 + \sqrt{(1 - 6\rho)(1 - 2\rho)}) > 1 - 3\rho$ when $\rho < 1/6$) by the Poincaré-Bendixson theorem. $\square$

10

Numerical analysis suggests that P_t is monotone decreasing with t; if so, then for $\rho < 1/6$ we would have: $P_t \geq \frac{1}{2}(1 + \sqrt{(1 - 6\rho)(1 - 2\rho)}) > 1 - 3\rho$ for all $t > 0$. This monotoniticy of P_t with t can be shown to be equivalent to the condition that $E_t \leq 2\rho$ for all $t \geq 0$, and so a formal proof of this inequality would be helpful.

It would also be interesting to obtain corresponding results to Theorem 2.3 for maximum parsimony applied to more general models - particularly for symmetric models on more than two states (some limited results are described in [18], Sections 9.4.1 and 9.5.1). Here we offer the following:

Conjecture 2.4. *For the r-state symmetric Poisson model Proposition 2.1 generalizes to give a lower bound on P^* of $1 - c_r \cdot l_+$ for some constant $c_r > 0$. Similarly, Theorem 2.3 generalizes to give an analogous result, where the critical ratio $\lambda/\mu = 6$ is replaced by $\lambda/\mu = c'_r$ for some constant $c'_r > 0$.*

3. Case II: Conservative GTR proceses

For any Markov process, we often write $\mathbb{P}_i(X_t = j)$ or $p_{ij}(t)$ for the conditional probability $\mathbb{P}(X_t = j | X_0 = i)$ that $X_t = j$ given that $X_0 = i$. We will say that a GTR model is *conservative* if, for every state i we have:

$$p_{ii}(t) > p_{ij}(t) \text{ for all } t \geq 0 \text{ and all } j \neq i.$$

This is the 'forward inequality' described by Sober [15]. The Kimura two-parameter (K2P) model (and every submodel, such as Jukes-Cantor) is an example of a conservative process (see Fig. 2). In this model the substitution probabilities are given as follow (for details see [20]):

$$p_{ij}(t) = \begin{cases} \frac{1}{4}(1 + e^{-\mu t} + 2e^{-\mu t(\kappa+1)/2}), & \text{if } i = j; \\ \frac{1}{4}(1 + e^{-\mu t} - 2e^{-\mu t(\kappa+1)/2}), & \text{if } i \to j \text{ is a transition}; \\ \frac{1}{4}(1 - e^{-\mu t}), & \text{if } i \to j \text{ is a transversion.} \end{cases}$$

With a conservative model, the majority rule can be consistent for ancestral reconstruction. Assuming state i at the tree root, the probability of
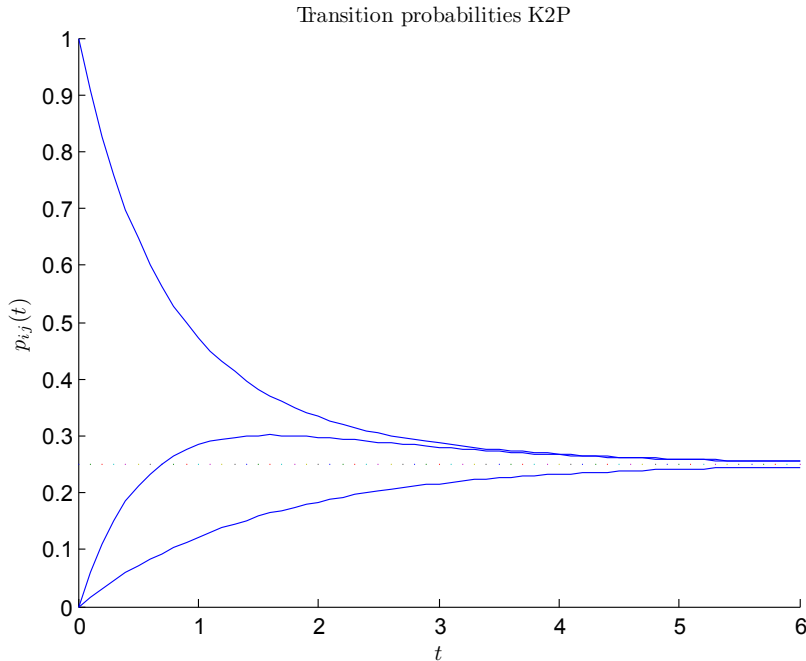
Figure 2: The three substitution probabilities for the Kimura 2ST model. This model is conservative (but not monotone). In this example, $\kappa = 4$ and μ is chosen to be 2/3 so that t corresponds to the expected number of substitutions. The curve descending from 1 is the function $p_{ii}(t)$. The middle curve, which has a local maximum around 1.6 is the probability of a transition (A $\leftrightarrow$ G or C $\leftrightarrow$ T); the lower curve is the probability of a transversion (Purine (A or G) $\leftrightarrow$ Pyrimidine (C or T).

observing i at any tree leaf is higher than the probability of observing any particular alternative state j. This holds true for whatever the root-to-leaf distances and the tree topology. With a star tree, with an upper bound on the root-to-leaf distances, the probability that this inference rule makes the correct selection tends to 1 as the number of leaves grow (by the central limit theorem for sums of independent random variables). We shall see that this result still holds for a more general class of trees under mild assumptions. We now describe this class of trees and their properties.

3.1. Well-spread trees

Given a rooted phylogenetic X-tree and a leaf x in the leaf set X, let:

$$l_x := \sum_{e \in P(\rho,x)} l_e,$$

the sum of the branch lengths on the path $P(\rho, x)$ from the root of the tree (ρ) to leaf x. Similarly, for distinct leaves $x, y \in X$, let:

$$l_{xy} = \sum_{e \in P(\rho,x) \cap P(\rho,y)} l_e,$$

the total length of the shared path from ρ to the leaves x, y. Finally, define the *spread* of T as:

$$s(T) := \frac{\sum_{x,y} \min\{l_{xy}, 1\}}{n(n-1)}.$$

Thus, provided $l_{xy} < 1$ for all x, y, $s(T)$ is the average value of l_{xy} over pairs x, y. We say that T is *well spread* if $s(T)$ is small; more precisely, T is $1 - \beta$ spread if $s(T) \leq \beta$. In particular, a tree is a star tree if and only if it is 1-spread (since 1-spread is equivalent to the condition $s(T) = 0$ which in turn is equivalent to the condition $l_{xy} = 0$ for all distinct pairs x, y and this holds precisely if and only if the tree is a star tree).

Note that a well-spread tree must have a large number of edges close to its root; an example is shown in Fig. 3.
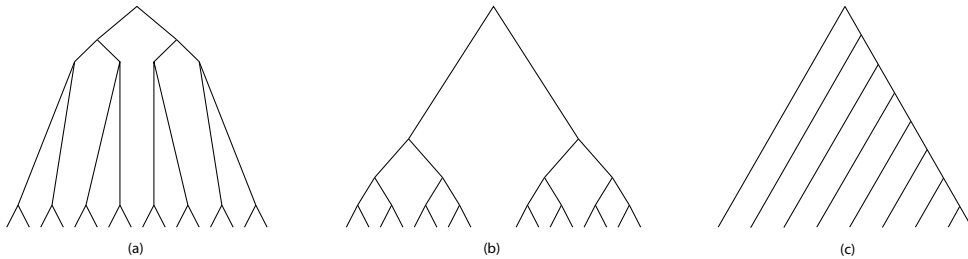

Figure 3: (a) A well-spread tree; (b, c) Trees that are not well-spread.

It is easily shown that a sufficient condition for a tree to be $1 - \beta$ spread is that, for some $\epsilon, \delta > 0$ with $\epsilon + \delta < \beta$, the proportion of pairs of leaves

whose paths from the root to those leaves overlap by a length of at least ϵ is no more than δ. We use this observation to show that the spread of a Yule pure-birth tree of fixed height t approaches 1 as the speciation rate grows.

Proposition 3.1. *Consider a random Yule pure-birth tree T that has speciation rate λ and fixed height t. Then for any $\beta > 0$, the probability that T is $1 - \beta$ spread converges to 1 as $\lambda \to \infty$.*

Proof. We may assume that T has a root of out-degree 2 (the length of a single lineage shrinks to zero with probability 1 as λ grows). By Theorem 2(2) of [10], the expected proportion of pairs of leaves whose most common ancestor lies r or more edges from the root of the tree has the geometric probability $(2/3)^r$. Given $\epsilon, \delta > 0$ with $\epsilon + \delta < \beta$ first select a sufficiently large value of r that $(2/3)^r \leq \delta$. For any $\eta > 0$ we can now select a sufficient large value of λ that the probability that all the (at most) 2^r vertices separated from the root by r edges have are within distance ϵ from the root is at least $1 - \eta$. The result now follows. $\square$

We now introduce some further notation. For each state $j \in \mathcal{S}$, let n_j denote the number of leaves of T that are in state j, and let ρ_j^i be the expected proportion of leaves that are in state j, given that the root is in state i. Thus n_j is a random variable (whose distribution depends on the root state i) while ρ_j^i is a value determined by the model parameters, j and root state i.

The following Lemma is central to many of the results that follow in this paper.

Lemma 3.2. *Suppose that T is a rooted tree, with branch lengths, which is $1 - \beta$ spread. Then for any continuous-time Markov process on T, the following holds for all initial states i: For any $s > 0$, the probability of the event that for all states $j \in \mathcal{S}$:*

$$\left| \frac{n_j}{n} - \rho_j^i \right| < s$$

is at least $1 - f(n, \beta)/s^2$ where $f(n, \beta)$ tends to 0 as $\max\{\frac{1}{n}, \beta\} \to 0$.

Proof. For $x \in X = \{1, \ldots, n\}$, let θ_x^j be the random variable that takes the value 1 if leaf x is in state j and 0 otherwise. We have $\frac{n_j}{n} = \frac{1}{n}\sum_{x=1}^{n}\theta_x^j$ and $\rho_j^i = \frac{1}{n}\sum_{x=1}^{n}p_{ij}(l_x)$. In particular, since $p_{ij}(l_x) = \mathbb{E}[\theta_x^j]$, linearity of expectation gives:

$$\mathbb{E}\left[\frac{n_j}{n}\right] = \rho_j^i.$$

Now:

$$\mathrm{Var}\left[\frac{n_j}{n}\right] = n^{-2}\left(\sum_{x \in X}\mathrm{Var}[\theta_x^j] + \sum_{x,y \in X, x \neq y}\mathrm{Cov}[\theta_x^j, \theta_y^j]\right), \qquad (6)$$

and $\mathrm{Var}[\theta_x^j] \leq \frac{1}{4}$, $|\mathrm{Cov}(\theta_x^j, \theta_y^j)| \leq 1$. Moreover, for any pair (x,y) we claim that $|\mathrm{Cov}(\theta_x^j, \theta_y^j)| \leq K\min\{1, l_{xy}\}$ for a constant K dependent only on the model. To see this, let N_{xy} be the event that the root ancestral state (i) does not change state anywhere along the shared path of length l_{xy}. We have $\mathbb{P}(N_{xy}) = \exp(-cl_{xy})$ for a constant c that is dependent only on the model and the ancestral state i. Moreover, the random variables θ_x^j, θ_y^j are conditionally independent, given N_{xy}. Routine algebra then shows that we can express $\mathrm{Cov}[\theta_x^j, \theta_y^j]$ as l_{xy} times a constant, plus terms of order l_{xy}^2. However, since in addition $\mathrm{Cov}[\theta_x^j, \theta_y^j] \leq 1$, we have $\mathrm{Cov}[\theta_x^j, \theta_y^j] \leq \min\{Kl_{xy}, 1\} \leq K\min\{l_{xy}, 1\}$ for some sufficiently large constant $K > 1$. Thus, from 6, we have:

$$\mathrm{Var}\left[\frac{n_j}{n}\right] \leq \left(\frac{1}{4n} + Ks(T)\right).$$

Let $f_1(n, \beta) = (\frac{1}{4n} + K\beta)$ then, since $s(T) < \beta$, and by Chebyshev's inequality, we have:

$$\mathbb{P}\left(|\frac{n_j}{n} - \rho_j^i| \geq s\right) \leq \frac{\mathrm{Var}\left[\frac{n_j}{n}\right]}{s^2} \leq f_1(n, \beta)/s^2.$$

Thus, if $r = |\mathcal{S}|$ denotes the number of possible states, and if we let $f(n, \beta) := rf_1(n, \beta)$ then we have:

$$\mathbb{P}\left(\exists j : |\frac{n_j}{n} - \rho_j^i| \geq s\right) \leq f(n, \beta)/s^2,$$

which converges to zero when both $n \to \infty$ and $\beta \to 0$. $\qquad\square$

Theorem 3.3. *For a conservative model and a $1 - \beta$ spread tree, with $l_x \leq l$ for each x, the probability that the majority state at the leaves is identical to the ancestral state at the root is at least $1 - g(n, \beta, l)$ for a function g which (for each value l) tends to zero as $\max\{\frac{1}{n}, \beta\} \to 0$.*

15

Proof. Let
$$\delta_l := \min_{i,j:i\neq j} \inf\{p_{ii}(t) - p_{ij}(t) : t \in [0,l]\}.$$

Since p is continuous, and $[0,l]$ is compact, the conservative property implies that $\delta_l > 0$. Moreover, we have:

$$\rho_i^i - \rho_j^i \geq \delta_l \text{ for all } j \neq i. \qquad (7)$$

Now by Lemma 3.2, the probability of the event that $\left|\frac{n_j}{n} - \rho_j^i\right| < \frac{1}{2}\delta_l$ for all j is at least $1 - 4f(n,\beta)/\delta_l^2$. Moreover, for this event, Inequality (7) implies (by the triangle inequality) that $\frac{n_i}{n} - \frac{n_j}{n} > 0$ for all $j \neq i$; that is, the ancestral state i is the majority state at the leaves. Thus the probability that the ancestral state is the majority state is at least $1 - g(l,n,\beta)$ where $g(l,n,\beta) := 4f(n,\beta)/\delta_l^2$ has the required stated properties. $\qquad\square$

4. Case III: Monotone time-reversible proceses

Note that, for *any* general time-reversible (GTR) Markov process the function $p_{ii}(t)$ is always monotone decreasing to its equilibrium frequency π_i for each state i [2], that is:

$$p_{ii}(t) > p_{ii}(t') \text{ for all } t < t'.$$

We will say the model is *monotone* if, for all distinct states i,j, we have:

$$p_{ij}(t) < p_{ij}(t') \text{ for all } t < t'.$$

Thus a monotone model has the property that if we start in a particular state i then the probability that we are in a different particular state j at time t increases monotonically with t towards its equilibrium probability π_j. In particular, a monotone GTR model satisfies the 'backward inequality' from [15] that $p_{ii}(t) > p_{ji}(t)$ for all $j \neq i$ and $t \geq 0$, since $p_{ji}(t)$ is monotone increasing to π_i while $p_{ii}(t)$ is monotone decreasing to π_i.

For any number of states, models such as the Felsenstein 1981 model (also called the F81, Tajima-Nei, or Equal Input model) are monotone (but not conservative, unless all equilibrium frequencies are equal). Also any two–state Markov process is monotone (Fig. 4) and the implications of this for
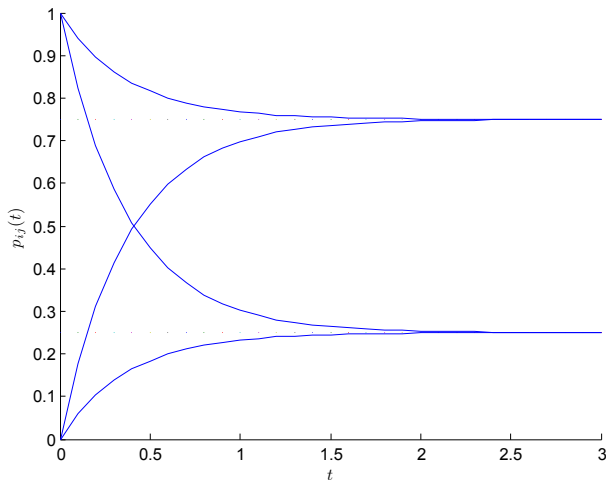
16

Figure 4: A two-state model with different equilibrium frequencies $(0.75, 0.25)$ for the states 0 and 1, respectively. The two decreasing curves are $p_{00}(t)$ (upper) and $p_{11}(t)$ (lower). The two increasing curves are $p_{10}(t)$ (upper) and $p_{01}(t)$ (lower). This model is monotone, but not conservative.

biological inference on the basis of a single observation ($n = 1$) were explored in [15] and [16].

Amongst nucleotide substitution models, the K2P model is not monotone, since if $i \neq j$ represents a transition then:

$$p_{ij}(t) = \frac{1}{4} + \frac{1}{4}e^{-\mu t} - \frac{1}{2}e^{-\mu t(\frac{\kappa+1}{2})}$$

can behave as shown by the middle curve in Fig. 2, where κ is the transition-transversion ratio (taken to be a default option of 4 here).

Despite K2P not being monotone, this model nevertheless satisfies Sober's 'backward inequality' as it is a symmetric model (i.e. $p_{ij}(t) = p_{ji}(t)$ for all t); however more complex time-reversible continuous Markov processes can fail this inequality. For example, consider a process on states $0, 1, 2, \ldots, m$ with equal and high transition rates from each value of k (less than m) to $k + 1$ and equal low transition rates from each k (greater than 0) to $k - 1$. Then for a suitably large value of m and choice of $t = t_1$, we have $p_{11}(t_1) < p_{01}(t_1)$. In particular, observing state 1 at a particular (known) time t_1 provides more evidence that the initial state was 0 rather than 1.

17

With monotone models, the majority rule can be misleading – when the time t is larger than the time corresponding to the intersection point $p_{ii}(t) = p_{ij}(t)$, it is more likely to have $j \neq i$ at any given leaf than to have i. However, simple prediction rules still exist, which depend on whether the equilibrium frequencies are known.

When the equilibrium frequencies are known, we use the fact that if i is the ancestral state then the proportion of taxa in state i, $\frac{n_i}{n}$ is expected to be larger than π_i (at least if the number of taxa is sufficient to avoid sampling effects), while $\frac{n_j}{n}$ is expected to be less than π_j for all $j \neq i$. This suggests a *modified majority rule*: select as an ancestral state estimate the state i which maximizes $\frac{n_i}{n} - \pi_i$. Note that the branch lengths and even the tree topology do not need to be known. Moreover, this decision rule becomes the simple majority rule when the equilibrium frequencies are all equal.

However, with site specific models, we cannot assume that the equilibrium frequencies are known, especially with proteins (as discussed earlier). In such a case, we can use a second decision rule, based on the fact that $p_{ii}(t)$ is a decreasing function of t while $p_{ij}(t)$ is increasing. This second decision rule needs the root-to-leaf distances to be known and variable across taxa (however the tree topology may be unknown). Let $\overline{l_j}$ be the average distance between the root and the taxa having state j, and let $\overline{l_{-j}}$ be the average distance between the root and the taxa having a state different to j. As the distance of a leaf from the root increases, the probability that leaf is in the ancestral state i should also decrease, while the reverse trend should hold for any other state j. In other words, we select i to minimize $\overline{l_i} - \overline{l_{-i}}$. Note that for this rule to apply, we need the root-to-taxon distance to be sufficiently heterogeneous. With a molecular clock-tree this rule is of no help. Moreover, we do not need to know the site rates and the absolute branch lengths, and the topology and branch lengths may be unknown, provided we still can estimate the root-to-leaf distances.

We shall see that under mild assumptions, both rules for monotone models are statistically consistent. We now describe the two procedures for monotone models more precisely, depending on whether π is known or not. We then state a theorem that provides conditions under which these estimators are accurate.

The two procedures are as follows:

- π **known:** Select the ancestral state i to maximize $\frac{n_i}{n} - \pi_i$.

- π **not known:** Select the ancestral state i to minimize $\overline{l_i} - \overline{l_{-i}}$.

We will show that the first estimator performs well provided the tree is well spread and n is large. The second estimator requires, in addition, that there be reasonable spread amongst the root-to leaf distances (i.e. that the tree *not* be clocklike). First, we require a lemma which is a mild extension of Chebyshev's order inequality (the proof is given in the Appendix).

Lemma 4.1. *Suppose that Y is a random variable taking values in $[0, l]$ and that $f : [0, l] \to \mathbb{R}$ is a smooth function with $f'(y) \geq c > 0$ for all $y \in [0, l]$. Then:*
$$\text{Cov}[Y, f(Y)] \geq c \cdot \text{Var}[Y].$$

Similarly, if $f'(y) \leq -c < 0$ for all $y \in [0, l]$ then $\text{Cov}[Y, f(Y)] \leq -c \cdot \text{Var}[Y]$.

Theorem 4.2. *Suppose we have a monotone GTR model and $\alpha > 0$.*

1. *The first estimation procedure described above (for a known π) correctly selects the true ancestral state with probability at least $1 - \alpha$ provided the following three conditions hold:*

   (i) $l_x \leq l < \infty$, for all x, and some l independent of n;

   (ii) T is $1 - \beta$ spread for sufficiently small values of β, and

   (iii) n is sufficiently large.

2. *The second estimation procedure described above (for π not known) correctly selects the true ancestral state with probability at least $1 - \alpha$ provided that, in addition to conditions (i) – (iii), the following two conditions hold:*

   (iv) The variance of the l_x values is greater or equal to some fixed value $v > 0$ that is independent of n.

   (v) $\pi_j \neq 0, 1$, for any $j \in \mathcal{S}$.

19

Proof. For part (1), let $\delta_1 = \min_{i,j:i\neq j}\{\pi_j - p_{ij}(l)\}, \delta_2 = \min_i\{p_{ii}(l) - \pi_i\}$ and $\delta_l = \min\{\delta_1, \delta_2\}$. By the monotonicity property, we have $\delta_l > 0$. If i is the ancestral state then:

$$\rho_i^i \geq \pi_i + \delta_l, \text{ and for any state } j \neq i, \rho_j^i \leq \pi_j - \delta_l. \tag{8}$$

Now by Lemma 3.2, the probability of the event that $\left|\frac{n_j}{n} - \rho_j^i\right| < \delta_l$ for all j is at least $1 - f(n,\beta)/\delta_l^2$. Moreover, for this event, Inequality (8) implies (by the triangle inequality) that $\frac{n_i}{n} - \pi_i > 0$ and for all $j \neq i$, we have $\frac{n_j}{n} - \pi_j < 0$, in which case the correct ancestral state (i) will be selected by the decision rule. Thus if we select a sufficiently small value of β and a sufficiently large value of n that $1 - f(n,\beta)/\delta_l^2 < \alpha$ we obtain the result in Part (1).

For part (2), we show that if i is the ancestral state then, with high probability, $\overline{l_i} - \overline{l_{-i}} < 0$ and for all $j \neq i$, $\overline{l_j} - \overline{l_{-j}} > 0$. For any state j (including i), consider the difference:

$$D_j := \overline{l_j} - \overline{l_{-j}}.$$

Recalling the definition of θ_x^j from the proof of Lemma 3.2 we have:

$$\overline{l_j} = \frac{\sum_{x\in X} l_x \theta_x^j}{n_j} \text{ and } \overline{l_{-j}} = \frac{\sum_{x\in X} l_x(1 - \theta_x^j)}{(n - n_j)},$$

and so:

$$D_j = \frac{\frac{1}{n}\sum_{x\in X} l_x\theta_x^j - (\frac{1}{n}\sum_{x\in X} l_x)\cdot\frac{n_j}{n}}{\frac{n_j}{n}(1 - \frac{n_j}{n})}. \tag{9}$$

By assumption (v), D_j is well defined (i.e. $n > n_j > 0$ in the denominator) with probability converging to 1 as n grows. Let

$$\overline{l} := \frac{1}{n}\sum_{x\in X} l_x, \text{ and let } \overline{L} := \frac{1}{n}\sum_{x\in X} l_x p_{ij}(l_x).$$

Notice that we can write the numerator of D_j in the form:

$$(\overline{L} - \overline{l}\rho_j^i) + \left(\frac{1}{n}\sum_{x\in X} l_x\theta_x^j - \overline{L}\right) + \overline{l}(\rho_j^i - \frac{n_j}{n}). \tag{10}$$

Now, let $c_1 = \min_{j\neq i}\inf\{\frac{dp_{ij}(t)}{dt} : t \in [0,l]\}$ and $c_2 = \inf\{\frac{-dp_{ii}(t)}{dt} : t \in [0,l]\}$, and $c = \min\{c_1, c_2\}$. By the monotone assumption, $c > 0$. We can now apply

Lemma 4.1 as follows. Define a random variable Y by setting $Y = l_x$ for a leaf x selected uniformly at random from the leaf set X, and let $f(y) = p_{ij}(y)$. Then, $\mathrm{Cov}[Y, f(Y)] = \overline{L} - \overline{l}\rho_j^i$ and so, by Lemma 4.1, we have:

$$\overline{L} - \overline{l}\rho_i^i \leq -cv, \text{ and } \overline{L} - \overline{l}\rho_j^i \geq cv \text{ for all } j \neq i, \tag{11}$$

where $v > 0$ is a lower bound on the variance of the l_x values from condition (iv). Note that $\mathbb{E}[\frac{1}{n}\sum_{x \in X} l_x \theta_x^j] = \overline{L}$, and since $l_x \leq l$ for all $x \in X$, an argument similar to that given in Lemma 3.2 implies that $\left|\frac{1}{n}\sum_{x \in X} l_x \theta_x^j - \overline{L}\right|$ can be made less than any $\delta > 0$ by selecting β and $\frac{1}{n}$ sufficiently small. Moreover, the same applies for the difference $\left|\frac{n_j}{n} - \rho_j^i\right|$ by Lemma 3.2. Thus, from expression (10), the numerator of D_j can be made arbitrarily close to the difference $\overline{L} - \overline{l}\rho_j^i$ by selecting β and $\frac{1}{n}$ sufficiently small. It then follows from Inequality (11) that the sign of D_j will be negative for $j = i$ and positive otherwise, as required (noting that c depends just on the model, not on β or n). This completes the proof.

$\square$

5. Case IV: Non-monotone and non-conservative models

Some simple and widely used models are neither monotone nor conservative. For example, the 'HKY' (Hasegawa, Kishino and Yano) model combines both K2P and F81 ([20]); as with K2P, the transition probabilities first increase and then decrease (non-monotony); because the equilibrium frequencies may be unequal, the probability of observing the root state i at a leaf may be less than the probability of observing state j when $\pi_i < \pi_j$.

With such models, the justifications provided for the statistical consistency of the three simple rules above and parsimony no longer apply. However, when the model is fully known and the tree is clock-like, the ancestral state can still be estimated using the frequencies of the states at the tree leaves. We shall see that this method is statistically consistent.

We first state a general result concerning general Markov processes.

Lemma 5.1. *Consider any continuous-time, irreducible Markov process, and let X_t be the state at time t. Then for any given $t \geq 0$, the probability*
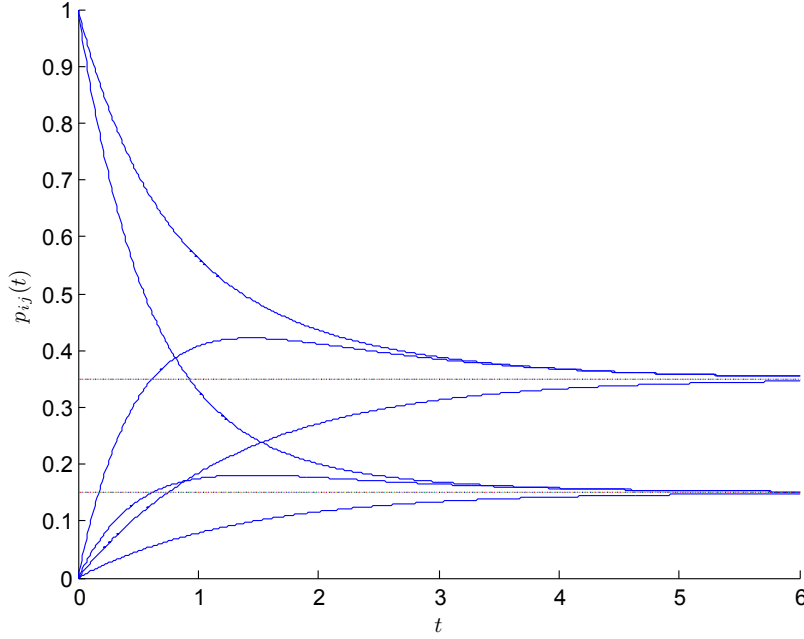
Figure 5: HKY transition probabilities with standard parameter values ($\kappa = 4$, purine=pyrimidine=0.5, G+C = 70%). The asymptotic values are the equilibrium frequencies of G and C (0.35) and A and T (0.15). The two decreasing curves are $p_{ii}(t)$ (e.g. $G \to G$ and $C \to C$ for the top-most curve). The two increasing curves with local maxima are for transitions (e.g. $A \to G$, $T \to C$ for the top-most increasing curve) while the two monotone increasing curves are for transversions.

distribution on X_t determines both X_0 and t. That is:

$$p_{ij}(t) = p_{i'j}(t') \text{ for all } j \in \mathcal{S} \Rightarrow i = i', t = t'.$$

Proof. Let $\mathbf{e}_i$ be the vector that has 1 in position i and 0 otherwise, and define $\mathbf{e}_{i'}$ analogously. Now, the vector $\mathbf{p}^i(t) := [p_{ij}(t) : j \in \mathcal{S}]$ satisfies $p^i(t) = \mathbf{e}_i \cdot \exp(Qt)$; similarly we have $\mathbf{p}^{i'}(t') = \mathbf{e}_{i'} \cdot \exp(Qt')$. Suppose values of t, t' exist for which $\mathbf{p}^i(t) = \mathbf{p}^{i'}(t')$. Without loss of generality, we may suppose that $t \geq t'$. In this case we have:

$$(\mathbf{e}_i \cdot \exp(Q(t - t')) - \mathbf{e}_{i'}) \cdot \exp(Qt') = \mathbf{0}.$$

Moreover, since the process is irreducible, $\mathbf{e}_i \cdot \exp(Q(t-t'))$ can equal $\mathbf{e}_{i'}$ only if $t = t'$ and $i = i'$, so if this is not the case, we have $\mathbf{w} \cdot \exp(Qt') = \mathbf{0}$ for a

22

non-zero vector **w** which implies that

$$\det \exp(Qt') = 0.$$

But, by Jacobi's identity, $\det \exp(Qt') = \exp(tr(Q)t') > 0$. This completes the proof. $\square$

From this Lemma, it follows that for the very special case of a star tree with all edges of equal length we can use maximum likelihood to consistently infer the ancestral state. This is because, in this very special case, the states at the n leaves provide n i.i.d. samples of the process, and so the identifiability conditions required to estimate both the ancestral state s_0 and the common edge length simultaneously hold (for similar reasons to the tailored argument for the consistency of MLE in settings such as phylogenetic tree reconstruction, described in Lemma 5.1 of [3]). Moving from star trees to the more general class of well-spread trees, we have the following main result of this section:

Theorem 5.2. *Suppose we have a continuous-time irreducible Markov process with rate matrix Q given, and let $\alpha > 0$. Consider a rooted phylogenetic tree on n leaves, for which the branch lengths l_e satisfy a molecular clock, i.e. $l_x = l_0$ for all x, where l_0 is less than some known value l. Assume also that the tree is $1 - \beta$ spread. Then we can estimate the ancestral state s_0 correctly with probability at least $1 - \alpha$ provided that n is sufficiently large, and β is sufficiently small.*

Proof. We will establish this result by a procedure that selects the state i for which the entire probability distribution $p_{ij}(t)$ (as j varies) can be made the 'closest' to the empirical distribution $\frac{n_j}{n}$ for an optimal value of t. We will use the l_∞ metric to measure 'closeness' (although, in applications other metrics may be preferable) so we will select the ancestral estimate i if i minimizes the quantity:

$$\inf_{t \in [0,l]} \max_j \left| \frac{n_j}{n} - p_{ij}(t) \right|.$$

First observe that, for any two states $i,' \in \mathcal{S}$ with $i \neq i'$, if we let:

$$\Delta_{ii'} := \inf_{t,t' \in [0,l]} \max_{j \in \mathcal{S}} \{|p_{ij}(t) - p_{i'j}(t')|\},$$

23

then $\Delta_{ii'} > 0$ by Lemma 5.1, the compactness of $[0, l]$ and the continuity of p. Thus $\delta_l := \min_{i,i':i\neq i'} \Delta_{i,i'}$ is also strictly greater than zero. Notice that δ_l is independent of n. Suppose that i is the true ancestral state and i' is a different state. By the molecular clock assumption, $\rho_j^i = p_{ij}(l_0)$. Thus, by Lemma 3.2, the probability of the event that $\left|\frac{n_j}{n} - p_{ij}(l_0)\right| < \frac{1}{2}\delta_l$ for all j is at least $1 - 4f(n, \beta)/\delta_l^2$. Moreover, for this event:

$$\inf_{t\in[0,l]} \max_j \left|\frac{n_j}{n} - p_{ij}(t)\right| < \inf_{t'\in[0,l]} \max_j \left|\frac{n_j}{n} - p_{i'j}(t')\right|$$

since the left-hand side is less than $\frac{1}{2}\delta_l$ and if t' is the value that minimizes the right-hand side then, by the triangle inequality for the l_∞ metric:

$$\max_j \left|\frac{n_j}{n} - p_{i'j}(t')\right| \geq \max_j |p_{ij}(t) - p_{i'j}(t')| - \max_j \left|\frac{n_j}{n} - p_{ij}(t)\right| \geq \Delta_{ii'} - \frac{1}{2}\delta_l \geq \frac{1}{2}\delta_l.$$

Thus, the selection method will choose the correct ancestral state (i) with probability at least $1 - 4f(n, \beta)/\delta_l^2$ and, as before, this can be larger than $1 - \alpha$ by ensuring that $\frac{1}{n}$ and β are sufficiently small. $\square$

When the tree is non-clock like, and the model and branch lengths are known, we might use a standard ML approach based on the pruning algorithm, though the precise conditions required for statistical consistency seem less clear.

6. Simulations

To compare the convergence rate and the performance of the various ancestral character reconstruction methods discussed in the previous sections, we performed computer simulations under biologically realistic conditions similar to [4].

We first generated a Yule tree with $n = 25$, 50, 100, 200, 400, 800 and 1600 leaves. This molecular-clock tree was then perturbed by multiplying every branch length (independently) by $(1 + X)$, where X was an exponential variable with parameter 0.5. The factor $(1+X)$ was used (as opposed to, say, X) to avoid an excessive number of very small branches. The observed departure from the molecular clock, as measured by the ratio between the longest

and shortest root-to-leaf lineages, was equal to ≈ 3.5 on average, a value that is usual in published phylogenies. Finally, the whole tree was re-scaled so that the average root-to-leaf distance was uniformly distributed between 0.1 (relatively low divergence) and 1.0 (high divergence). We generated 500 trees using this procedure (Yule process, molecular clock perturbation and re-scaling) for each tree size n.

DNA-like sequences of 100 sites were evolved along these trees using the HKY model with $\kappa = 4.0$ (default value in most software) and the equilibrium frequencies of A, C, G and T being equal to 0.15, 0.35, 0.35 and 0.15, respectively (such G+C bias is observed in thermophilic bacteria and archaea, while *Plasmodium* species have an even stronger A+T bias). The same parameter values are used in Fig. 5. This HKY model was combined with a discrete gamma distribution of parameter 1.0 with six rate categories. We thus obtained 500 data sets of 100 sites for each tree size n.

Five ancestral character prediction methods were compared:

- 'Parsimony' (studied in Section 2);

- 'Majority' (studied in Section 3);

- 'Modified majority', when the equilibrium frequencies π are known (studied in Section 4, *cf.* part (1) of Theorem 4.2);

- 'Difference of average root-to-leaf distances', when the equilibrium frequencies π are unknown, but we know the root-to-leaf distances (studied in Section 4, *cf.* part (2) of Theorem 4.2);

- 'Presence', which involves drawing with equal probability one of the character states that are present at the tree leaves. Indeed, it frequently occurs (notably with small n) that not all four possible states are observed at the tree leaves. Moreover, all previous prediction methods never output a character state that is not seen at the tree leaves. The difficultly of the prediction problem thus depends on the number of extant character states. In the extreme case where we observe a unique extant character state, all tested methods achieve perfect predictions unless hidden convergent substitutions (still possible but rare in our experiments, see below). On the other hand, when the four character

states are observed the chance is $\frac{1}{4}$ to be correct by chance. 'Presence' is thus used to re-scale the performance of the various methods, depending on the number of extant character states, which (among other factors) impacts the hardness of the prediction problem.

All methods were run with perfect knowledge of the tree topology ('Parsimony'), equilibrium frequencies π ('Modified majority') or root-to-leaf distances ('Difference of average root-to-leaf distances'). For each method and each data set, we measured:

- The percentage of correct predictions;

- The rescaled percentage of correct predictions, using the results achieved by 'Presence'. Let P be the percentage of correct predictions of the given method, and R be the percentage of correct predictions of 'Presence'; the rescaled percentage of correct predictions is equal to $(P - R)/(1 - R)$ and measures the fraction of improvement brought by the given method compared to random predictions.

Results averaged over 500 data sets are reported in Table 1 for each tree size n. We see that:

- The percentage of correct predictions of 'Presence' clearly decreases when n increases (~ 0.5 and ~ 0.25 for $n = 25$ and $n = 1600$, respectively). A unique extant character state is observed for $\sim 20\%$ of the sites with $n = 25$, and for only $\sim 0.1\%$ with $n = 1600$. In such a case (unique extant character state) the percentage of correct predictions is larger than 0.99, indicating that hidden convergent substitutions are very rare. Overall, the percentage of correct random predictions is thus inversely proportional to the number of extant character states, and, in this respect, $n = 1600$ is twice harder than $n = 25$.

- The accuracy of all methods improves with increasing n. However, the re-scaled percentage of correct predictions is required to see this effect with 'Difference of average root-to-leaf distances', which is the method with the slowest convergence rate. Overall the convergence rate is relatively slow, as with 1600 taxa the best methods have $\sim 90\%$ of correct

26

predictions only. This could be due to several factors: (1) the prediction problem is harder with 1600 taxa than with 25 (see the results of 'Presence') ; (2) some of the methods (namely 'Parsimony' and 'Majority', see below) are likely inconsistent for the larger values of the average root-to-leaf distance; (3) the spread of trees converges slowly to 0 (a required condition for the consistency of 'Majority', 'Modified majority' and 'Difference of average root-to-leaf distances'), as the average spread with 25 taxa is ~ 0.14 while with 1600 taxa it still is ~ 0.07 (Table 1).

- Surprisingly, 'Parsimony' is slightly behind 'Majority' and 'Modified majority'. This finding is also observed with Jukes-Cantor model (results not shown), and thus cannot be attributed to the chosen substitution model (HKY); it is likely due to the fact that some of the simulated trees show a high divergence, a condition where 'Parsimony' tends to perform poorly (see Theorem 2.3).

- Both 'Majority' and 'Modified majority' are very close, while we expected the latter to be better because it makes use of the equilibrium frequencies π. The explanation is likely related to the fact that in our simulations the average root-to-leaf distance is uniformly distributed in the $0.1 - 1.0$ range, a condition where HKY is conservative in most cases (HKY is conservative up to $t \approx 0.8$, see Fig. 5) and thus 'Majority' is often consistent. However, we see a small superiority of 'Modified majority' with large n, when the estimations of the n_i/n frequencies become sufficiently reliable. Moreover, HKY is monotone and 'Modified majority' is consistent up to $t \approx 1.45$ (Fig. 5), a value that is larger than the consistency limit of 'Majority' (~ 0.8) and the maximum expected root-to-leaf distance (1.0).

- Finally, the performance of 'Difference of average root-to-leaf distances' is rather low, but there is a clear improvement with large n. This confirms that root-to-leaf distances bring substantial information, which could be combined with other standard approaches to enhance accuracy in difficult cases.

All together, the most surprising outcome of these simulations is the performance of the (very simple) 'Majority' approach. It must be emphasized

27

Table 1: Average accuracy with simulated data. For each method we provide the percentage of correct predictions and (within parentheses) the re-scaled percentage of correct predictions (see text for definition). 500 data sets with 100 sites each were used for each number of taxa (n). Abbreviations are Spread: average spread of trees, Mod. Majority: 'Modified majority', Diff. Aver. Dist.: 'Difference of average root-to-leaf distances'.

n	Spread	Parsimony	Majority	Mod. Majority	Diff. Aver. Dist.	Presence
25	0.141	0.820 (0.652)	0.832 (0.674)	0.824 (0.659)	0.609 (0.214)	0.499
50	0.128	0.841 (0.728)	0.852 (0.746)	0.846 (0.736)	0.570 (0.237)	0.433
100	0.115	0.853 (0.772)	0.863 (0.788)	0.860 (0.784)	0.539 (0.262)	0.371
200	0.104	0.864 (0.802)	0.870 (0.811)	0.871 (0.813)	0.521 (0.285)	0.326
400	0.092	0.873 (0.822)	0.880 (0.833)	0.886 (0.842)	0.522 (0.324)	0.289
800	0.080	0.885 (0.844)	0.885 (0.844)	0.896 (0.858)	0.537 (0.362)	0.270
1600	0.075	0.890 (0.852)	0.891 (0.853)	0.906 (0.873)	0.567 (0.410)	0.261

that 'Majority' does not use any additional knowledge (tree topology, root-to-leaf distances or equilibrium frequencies), meaning that the gap could be larger if the other methods (e.g. 'Parsimony') were used with only approximate knowledge (e.g. an inferred tree topology).

7. Discussion

In this paper, we have described and analysed five approaches for inferring ancestral root state in taxon-rich trees: maximum parsimony, simple majority rule, modified majority rule, difference of average root-to-leaf differences, and best-fit of expected distribution of leaf states to the empirical distribution. The methods are all relatively simple and easily implemented, and require different model (and tree) assumptions in order to justify their accuracy. They can be applied in settings where one does not have enough information to carry out a full maximum likelihood analysis using the usual pruning algorithm, and so may be more suitable for site-specific models, where the process of evolution is likely to vary in a partially unknown way from character to character.

The price one might expect to pay for a method that requires fewer assumptions or detailed knowledge of underlying parameters is lower accuracy. Nevertheless, we have described several results which show that these methods (particular to the type of model in question) can still return the correct

ancestral state provided that the number of taxa (n) is sufficiently large, and the tree is sufficiently well-spread. We have shown that for Yule trees with a high speciation rate (as a token for high taxon coverage), we expect a tree of fixed height to become increasingly well-spread as n grows. It is clear that some type of assumption on the spread of the tree is necessary to avoid having two long branches near the root and the majority of lineage splitting well away from the root, in which case accurate root state inference is not possible.

Except for maximum parsimony, the methods described do not use the tree topology explicitly, but rely on just the distribution of states at the leaves, and perhaps their distance from the root. Accordingly, such methods may be more robust to tree mis-specification. Of the class of models described monotone models are perhaps the most relevant for application, since most GTR models are likely to be monotone (and even conservative) when restricted to amounts of evolutionary change that are commonly encountered for sequence evolution.

Our choice of methods to study in this paper has been guided by what can be successfully analysed, and we are not advocating these methods above others that might be considered; in particular, we make no claim that they are 'best possible'. Indeed, if one has sufficient information then more standard approaches such as maximum likelihood would be preferable. However, the simplicity of these methods, and the fact that they are relatively robust to model mis-specification may make them a useful complement to more sophisticated approaches. For future studies, it would be worthwhile to explore the performance of these approaches on biological data-sets, comparing them with other alternative approaches that have been advocated. It would also be worth developing statistical tests to determine whether differences observed in the data by our approaches are significant or not.

8. Acknowledgments

References

[1] D. Aldous, Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today, Stat. Sci. 16 (2001), 23–34.

[2] D. Aldous, J.A. Fill, Reversible Markov chains and random walks on graphs, Chapter 3, Eqn. (40), 2010, http://www.stat.berkeley.edu/ aldous/RWG/book.html.

[3] J. Chang, Full reconstruction of Markov models on evolutionary trees: identifiability and consistency, Math. Biosci. 137 (1996), 51–73.

[4] R. Desper, O. Gascuel, Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting, Mol. Biol. Evol. 21(3) (2004), 587–98.

[5] W. Evans, C. Kenyon, Y. Peres, L.J. Schulman, Broadcasting on trees and the Ising Model, Ann. Appl. Probab. 10(2) (2000), 410–433.

[6] J.M. Koshi, R.A. Goldstein, Models of natural mutations including site heterogeneity, Proteins, 32 (1998), 289–295.

[7] N. Lartillot, H. Philippe, A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process, Mol. Biol. Evol. 21 (2004), 1095–1109.

[8] G. Li, J. Ma, L. Zhang, Greedy selection of species for ancestral state reconstruction on phylogenies: Elimination is better than insertion, PLoS ONE 5(2) e8985 (2010).

[9] D.A. Liberles, Ancestral sequence reconstruction, Oxford University Press, New York, 2007.

[10] A. McKenzie, M.A. Steel, Properties of phylogenetic trees generated by Yule-type speciation models, Math. Biosci. 170 (2001), 91–112.

[11] W.P. Maddison, Calculating the probability distributions of ancestral states reconstructed by parsimony on phylogenetic trees, Syst. Biol. 44 (1995), 474–481.

[12] E. Mossel, On the impossibility of reconstructing ancestral data and phylogenies, J. Comput. Biol. 10(5) (2003), 669–678.

[13] M. Pagel, The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies, Syst. Biol. 48 (1999), 612–622.

[14] B.A. Salisbury, J. Kim, Ancestral state estimation and taxon sampling density, Syst. Biol. 50(4) (2001), 557–564.

[15] E. Sober, Temporally asymmetric inference in a Markov Process, Phil. Sci. 58(3) (1991), 398–410.

[16] E. Sober, Evolution and Evidence: The logic behind the science, Cambridge University Press, Cambridge, UK, 2008.

[17] M.A. Steel Distributions on bicoloured evolutionary trees. PhD Thesis. Massey University, 1989.

[18] M. Steel, D. Penny, Maximum parsimony and the phylogenetic information in multi-state characters, in: V. Albert (Ed.), Parsimony, phylogeny and genomics, Oxford University Press, 2005, pp. 163–178.

[19] S.H. Strogatz, Nonlinear dynamics and chaos, Addison-Wesley Publishing Company, 1994.

[20] D.L. Swofford, G.J. Olsen, P.J. Waddell, D.M. Hillis, Phylogenetic Inference, in: D.M. Hillis, C. Moritz, B.K. Mable (Eds.), Molecular Systematics, Second Edition, Sinauer Press, 1996.

[21] C. Tuffley, M.A. Steel, Links between maximum likelihood and maximum parsimony under a simple model of site substitution, B. Math. Biol. 59(3) (1997), 581–607.

[22] J. Zhang, M. Nei, Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. J. Mol. Evol. 44 (1997), S139–S146.

[23] L.X. Zhang, J. Shen, J.L. Yang, G.L. Li, Analyzing the Fitch method for reconstructing ancestral states on ultrametric phylogenetic trees. Bull. Math. Biol, 2010, in press.

9. Appendix: Proof of Proposition 2.2 and Lemma 4.1

For the proof of Proposition 2.2 let $n = e^{\lambda t/2}$. Then the expected number of taxa at time $t/2$ is n and is n^2 at time t. Let N_u be the number of individuals at time u. Let E_1 be the event that $N_{\frac{t}{2}}$ lies between $\frac{1}{2}n$ and $\frac{3}{2}n$, let E_2 be the event that $N_t < 2n^2$, and let E be the conjunction of E_1, E_2. We first establish the following:

CLAIM: For some $\delta > 0$, $\mathbb{P}(E) \geq \delta$, for all sufficiently large λ.

We have $\mathbb{P}(E) = \mathbb{P}(E_2|E_1) \cdot \mathbb{P}(E_1)$. Now, the fact that $N_{t/2}/e^{\lambda t/2}$ has a limiting distribution as λ tends to infinity (an exponential distribution with a mean of 1) implies that $\mathbb{P}(E_1) \geq \delta' > 0$ for a fixed $\delta' > 0$ (we can take for δ' any number smaller than $e^{-\frac{1}{2}} - e^{-\frac{3}{2}}$ for large enough values of λ). Moreover, $\mathbb{E}[N_t|E_1] \leq \frac{3}{2}n^2$, since $\mathbb{E}[N_t|N_{t/2} = k] = ke^{\lambda t/2} = kn \leq \frac{3}{2}n^2$ for any $k \leq 3n/2$. However:

$$\mathbb{E}[N_t|E_1] \geq 2n^2 \cdot \mathbb{P}(N_t \geq 2n^2|E_1) = 2n^2(1 - \mathbb{P}(E_2|E_1)).$$

Thus, $\mathbb{P}(E_2|E_1) \geq \frac{1}{4}$, and so, $\mathbb{P}(E) \geq \frac{1}{4}\delta' =: \delta > 0$ as claimed.

Suppose the number of individuals at time $t/2$ is m; label them $1, 2, \ldots, m$. For individual i, let n_i be the number of descendants at time t. Thus $\sum_{i=1}^{m} n_i$ is the total number of individuals at time t. Now we use a well-known property of the (discrete) Yule distribution – for a binary tree with n_i leaves, the probability that the root is incident with a leaf is exactly $2/n_i$. Now individual $i \in \{1, \ldots, m\}$ is not the root of a binary tree, but if the binary tree below i has the property just described, then either the edge i lies on, or an edge in the binary tree below it, has a length of at least $t/4$. Also if $n_i \leq 2$ then once again we must have at least one edge with a length of at least $t/4$.

For any particular value of m that satisfies event E_1, let p be the probability that none of the m individuals gives rise in this way to an edge of length at least $t/4$. Then p is bounded above (by independence) as follows:

$$p \leq \prod_{i=1}^{m}(1 - \frac{2}{n_i}), \tag{12}$$

where the n_i values satisfy constraints implied by E:

$$\sum_{i=1}^{m} n_i \leq 2n^2, \text{ and } m \geq \frac{1}{2}n,$$

as well as our assumption $n_i \geq 2$ for all i. Maximizing the term on the right-hand side of (12) subject to the constraint $\sum_{i=1}^{m} n_i \leq 2n^2$, we have:

$$p \leq (1 - \frac{2m}{2n^2})^m \sim e^{-m^2/n^2} \leq e^{-0.25}.$$

Thus, with probability at least $\delta(1 - e^{-0.25})$ there is an edge in the Yule tree having length at least $t/4$. This completes the proof of Proposition 2.2.

Proof of Lemma 4.1. Suppose $f'(y) \geq c > 0$ for all $y \in [0, l]$ and that Y is discrete taking finite values $l \geq y_1 \geq y_2 \geq \cdots \geq y_n \geq 0$ (other cases are similar), and let $p(y) = \mathbb{P}(Y = y)$. Then evaluating the following double sum by expanding out terms gives us the identity:

$$\sum_{i,j}(y_i - y_j)(f(y_i) - f(y_j))p(y_i)p(y_j) = 2\text{Cov}[Y, f(Y)]. \qquad (13)$$

However we can also write this double sum in the form:

$$2\sum_{i,j:i>j}(y_i - y_j)(f(y_i) - f(y_j))p(y_i)p(y_j) \geq 2c\sum_{i,j:i>j}(y_i - y_j)^2 p(y_i)p(y_j), \quad (14)$$

where the inequality holds since, for $y_i \geq y_j$ the condition $f'(y) \geq c$ for all $y \in [0, l]$ implies that $f(y_i) - f(y_j) \geq c(y_i - y_j)$ by the mean value theorem. Now,

$$2c\sum_{i,j:i>j}(y_i - y_j)^2 p(y_i)p(y_j) = c\sum_{i,j}(y_i - y_j)^2 p(y_i)p(y_j) = 2c\text{Var}[Y].$$

Applying this to Eqns. (13) and (14) gives the result claimed.