Mathematical Biosciences 227 (2010) 125–135

Contents lists available at ScienceDirect

Mathematical Biosciences

journal homepage: www.elsevier.com/locate/mbs

Inferring ancestral sequences in taxon-rich phylogenies

Olivier Gascuel [a], Mike Steel [b,*]

[a] *Méthodes et Algorithmes pour la Bioinformatique, LIRMM, CNRS – Université de Montpellier, 161 rue Ada, 34392 Montpellier, France*
[b] *Allan Wilson Centre for Molecular Ecology and Evolution, University of Canterbury, Christchurch, New Zealand*

ARTICLE INFO

Article history:
Received 8 April 2010
Received in revised form 28 June 2010
Accepted 8 July 2010
Available online 11 July 2010

Keywords:
Ancestral sequence
Markov process
Parsimony
Majority rule

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.

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. [5,8,11–14,23]) 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

* Corresponding author. Tel.: +64 3 3667001; fax: +64 3 3642587.
E-mail addresses: gascuel@lirmm.fr (O. Gascuel), m.steel@math.canterbury.ac.nz, mathmomike@gmail.com (M. Steel).

0025-5564/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.mbs.2010.07.002

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 $\Pi = \mathrm{diag}(\pi)$ 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 the l_e values 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 $\max\left\{\frac{1}{n}, \xi\right\}$ 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 ($l_e \leqslant l < \infty$), 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).

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 branch lengths 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 [22]). 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* $l_+ < \frac{1}{2}\log\left(\frac{4}{3}\right)$, *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:*

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

Proof. When l_+ 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 [19], that:

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

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:

$$\frac{\sqrt{(1-4g)(1-8g)}}{2(1-2g)^2} \geqslant \frac{1}{2}(1-8g), \quad \text{and } g \leqslant l_+. \qquad \square$$

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.