Journal of Theoretical Biology 404 (2016) 126–142

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

A comparison of ancestral state reconstruction methods for quantitative characters

Manuela Royer-Carenzi*, Gilles Didier

Aix-Marseille Université, CNRS, Centrale Marseille, I2M, UMR 7373, 13453 Marseille, France

HIGHLIGHTS

- We compare 5 reconstruction methods on simulated evolution with directional term.
- We derive the ancestral state distributions under evolution models with trend.
- We prove that ML, REML and GLS methods infer the same ancestral states.
- We bound the reconstruction error and take into account its variance.
- No Brownian-based method performs well as the directional term increases.

ARTICLE INFO

Article history:
Received 7 August 2015
Received in revised form
9 May 2016
Accepted 20 May 2016
Available online 24 May 2016

Keywords:
Ancestral state reconstruction
Maximum likelihood
Brownian motion
Energy distance

ABSTRACT

Choosing an ancestral state reconstruction method among the alternatives available for quantitative characters may be puzzling. We present here a comparison of seven of them, namely the maximum likelihood, restricted maximum likelihood, generalized least squares under Brownian, Brownian-with-trend and Ornstein–Uhlenbeck models, phylogenetic independent contrasts and squared parsimony methods.

A review of the relations between these methods shows that the maximum likelihood, the restricted maximum likelihood and the generalized least squares under Brownian model infer the same ancestral states and can only be distinguished by the distributions accounting for the reconstruction uncertainty which they provide.

The respective accuracy of the methods is assessed over character evolution simulated under a Brownian motion with (and without) directional or stabilizing selection. We give the general form of ancestral state distributions conditioned on leaf states under the simulation models.

Ancestral distributions are used first, to give a theoretical lower bound of the expected reconstruction error, and second, to develop an original evaluation scheme which is more efficient than comparing the reconstructed and the simulated states.

Our simulations show that: (i) the distributions of the reconstruction uncertainty provided by the methods generally make sense (some more than others); (ii) it is essential to detect the presence of an evolutionary trend and to choose a reconstruction method accordingly; (iii) all the methods show good performances on characters under stabilizing selection; (iv) without trend or stabilizing selection, the maximum likelihood method is generally the most accurate.

1. Introduction

Besides being essential to understand the process of character evolution, ancestral state reconstruction plays an important role in the study of ecological diversification and comparative analysis. We focus here on quantitative characters, i.e. measured as continuous variables such as weight, size etc.

From a methodological point of view, ancestral state reconstruction is a challenging problem which has been addressed by several approaches. The general question can be stated as follows. Taking as inputs the phylogeny of a set of organisms (given as a tree with branch lengths) and their character states, a reconstruction method has to infer – as accurately as possible – the character states of the ancestral organisms. The reconstruction approaches fall into two major classes: methods based on the parsimony principle (Fitch, 1971; Swofford and Maddison, 1987; Maddison, 1991; Collins et al., 1994), whose general idea is to

* Corresponding author.
E-mail address: manuela.royer-carenzi@univ-amu.fr (M. Royer-Carenzi).

http://dx.doi.org/10.1016/j.jtbi.2016.05.029
0022-5193/© 2016 Elsevier Ltd. All rights reserved.

impute the missing values of the tree by minimizing the sum of distances between ancestors and their direct descendant characters, and methods based on stochastic models of character evolution, mainly Brownian motion for continuous traits (Schluter et al., 1997; Pagel, 1999b; Huelsenbeck and Ronquist, 2001; Nielsen, 2002). Several authors discuss the advantages of stochastic approaches over parsimonious ones (Schluter et al., 1997; Mooers and Schluter, 1999; Pagel, 1999b; Nielsen, 2002; Huelsenbeck et al., 2003). An important point is that stochastic approaches take into account divergence times (branch lengths) while parsimonious methods do not. Moreover, stochastic approaches may provide probability distributions of the reconstructed ancestral states, accounting for their uncertainty and which can be used to develop hypothesis testing and confidence intervals.

In our study, we focus on seven reconstruction methods, namely the maximum likelihood, restricted maximum likelihood, generalized least squares under Brownian, Brownian-with-trend and Ornstein–Uhlenbeck models, phylogenetic independent contrasts and squared parsimony methods. Before comparing their accuracy, we review the methods and their relationship to each other. It turns out that the first three ones reconstruct the same ancestral states. These three methods may still be distinguished, and to some extent compared, since they provide different probability distributions of their uncertainty. There are a few model-based approaches which do not rely on the Brownian assumption. For instance, in Hansen (1997), Martins and Hansen (1997), Pagel (1998), Pagel (1999a), authors consider ancestral states reconstructions under the assumption that the character follows either a Brownian motion with trend or an Ornstein–Uhlenbeck model, corresponding to a directional or a stabilizing selection respectively (Hansen and Martins, 1996). To our knowledge, the only available reconstruction approaches based on directional or stabilizing model are provided by the java program *COMPARE* which performs general least squares reconstructions according several models (Martins, 1995), by the computer package BayesTraits (Pagel et al., 2004; Pagel and Meade, 2013), which uses Markov chain Monte Carlo (MCMC) methods to infer ancestral states, and by the R-package *phytools*, which performs, through numerical optimization, maximum likelihood reconstructions under a Brownian motion with trend (Revell, 2012).

Evaluating the respective performances of these methods is a natural and important question. Works aiming at answering this question proceed by comparing the reconstructed states with reference “trusted” ones. Such reference values for ancestral states may be obtained either by considering fossil character states or by simulating, via a stochastic model, artificial evolution of the character and by keeping track of the ancestral states observed during simulations (Martins, 1999). Webster and Purvis (2002) and Oakley and Cunningham (2000) assess several reconstruction methods with regard to measurements on fossils. They both observe that the methods are confounded by an evolutionary trend toward increasing size.

Our comparison of the seven methods is based on artificial evolution simulated under Brownian motions with and without directional or stabilizing selection. The artificial evolution runs on the phylogenetic tree of Pleistocene planktic Foraminifera (Webster and Purvis, 2002). Besides the fact that we consider evolution models with directional or stabilizing selection, a noticeable difference with previous works is that the reconstructed states are compared with regard to the ancestral state distributions conditioned on the simulated leaves, rather than with the simulated ancestral states as it is done usually. Intuitively, in this way, we compare the reconstructed state with all the possible realizations of the evolution process with the given simulated leaf states. Moreover the ancestral distribution conditioned on the leaves does reflect the uncertainty inherent to the stochastic character of evolution as modeled in simulations. In particular, it allows us to determine a lower bound of the expected reconstruction error as well as the reconstructed state achieving this lower bound. This can be seen as a transposition of ideas of (Steel and Szekely, 1999) and (Royer-Carenzi et al., 2013).

Another motivation of this work is to assess the relevance of the distributions provided by the methods for the reconstruction uncertainty. These distributions are expected to provide a greater amount of information than single values for ancestral states (Schluter et al., 1997; Polly, 2001). Altogether with our new comparison scheme, we compare the conditional ancestral distributions given the leaves with the distributions provided by the methods. A distance between distributions, called the *Energy distance* offers us a consistent framework to compare both reconstructed states and reconstructed probability distributions, with ancestral state distributions conditioned on leaves (Szekely and Rizzo, 2013). The Energy distance is strongly related to the absolute bias.

Finally, we provide exact, matrix-based, implementations of Brownian-based methods which were formerly based on numerical optimization algorithms. Some of our R-scripts have been incorporated into the `reconstruct` function of the `ape` R-package since version 3.2 (Paradis et al., 2004, https://cran.r-project.org/web/packages/ape/index.html). We also provide matrix-based implementations of generalized least square (and equivalently maximum likelihood) reconstructions under Brownian motion with trend and Ornstein–Uhlenbeck models. Our R-scripts are available at https://github.com/gilles-didier/Reconstruction.git.

The rest of the paper is organized as follows. In Section 2, we present three standard models of quantitative character evolution. Section 3 briefly describes the reconstruction methods and shows how they are related. Section 4 is devoted to our assessment protocol. We provide the form of the ancestral distributions conditioned on the leaf states under the simulation. These ancestral distributions are next used to define our evaluation protocol and to give a lower bound of the expected reconstruction error. In its final version, the protocol is based on the Energy distance between probability distributions, both for assessing the reconstructed states and the distribution provided by the methods. The results of our simulations are finally presented and discussed in Section 5.

2. Models of evolution for quantitative characters

2.1. Phylogenetic trees – notations

In the standard ancestral character reconstruction problem, one assumes that the evolutionary history of the species is known and given as a rooted phylogenetic tree with branch lengths.

Our typical tree contains $n + 1$ nodes (including leaves), among which r are internal nodes (excluding the root). By convention, the nodes are indexed in the following way:

- index 0 for the root,
- indices 1 to r for the other internal nodes,
- indices $r + 1$ to n for the leaves.

The nodes are numbered in such a way that if a node j descends from a node i then $j > i$. We put $p(j)$ for the index of the direct ancestor of the node j, τ_j for the length of the branch leading to j and T_j for the sum of the branch lengths between the root and j. Being given two nodes i and j, we put $m(i, j)$ for the index of their most recent common ancestor (mrca).

Let X be a random variable. We write f_X for its density function and $\mathbb{E}(X)$ for its expectation.