Received: 2 December 2016 | Revised: 21 January 2017 | Accepted: 28 January 2017

DOI: 10.1002/ece3.2837

ORIGINAL RESEARCH

WILEY Ecology and Evolution Open Access

Rapid maximum likelihood ancestral state reconstruction of continuous characters: A rerooting-free algorithm

Eric W. Goolsby

Department of Ecology and Evolutionary Biology, Brown University, Providence, RI, USA

Correspondence
Eric W. Goolsby, Department of Ecology and Evolutionary Biology, Brown University, Providence, RI, USA.
Email: eric.goolsby.evolution@gmail.com

Funding information
National Science Foundation, Grant/Award Number: DEB-1501215

Abstract

Ancestral state reconstruction is a method used to study the evolutionary trajectories of quantitative characters on phylogenies. Although efficient methods for univariate ancestral state reconstruction under a Brownian motion model have been described for at least 25 years, to date no generalization has been described to allow more complex evolutionary models, such as multivariate trait evolution, non-Brownian models, missing data, and within-species variation. Furthermore, even for simple univariate Brownian motion models, most phylogenetic comparative R packages compute ancestral states via inefficient tree rerooting and full tree traversals at each tree node, making ancestral state reconstruction extremely time-consuming for large phylogenies. Here, a computationally efficient method for fast maximum likelihood ancestral state reconstruction of continuous characters is described. The algorithm has linear complexity relative to the number of species and outperforms the fastest existing R implementations by several orders of magnitude. The described algorithm is capable of performing ancestral state reconstruction on a 1,000,000-species phylogeny in fewer than 2 s using a standard laptop, whereas the next fastest R implementation would take several days to complete. The method is generalizable to more complex evolutionary models, such as phylogenetic regression, within-species variation, non-Brownian evolutionary models, and multivariate trait evolution. Because this method enables fast repeated computations on phylogenies of virtually any size, implementation of the described algorithm can drastically alleviate the computational burden of many otherwise prohibitively time-consuming tasks requiring reconstruction of ancestral states, such as phylogenetic imputation of missing data, bootstrapping procedures, Expectation-Maximization algorithms, and Bayesian estimation. The described ancestral state reconstruction algorithm is implemented in the *Rphylopars* functions *anc.recon* and *phylopars*.

KEYWORDS

ancestral state reconstruction, fast methods, linear-time algorithm, maximum likelihood, phylogenetic comparative method, phylogenetic generalized least squares

1 | INTRODUCTION

Phylogenetic comparative methods provide a framework for studying phenotypic evolution across species while accounting for statistical nonindependence due to common descent (Felsenstein, 1985; Martins & Hansen, 1997). Ancestral state reconstruction offers a powerful context for studying evolutionary trajectories, such as the number of times a particular phenotype evolved, estimating the approximate timing of major evolutionary events, and inferring missing phenotypic values corresponding to discovered fossils (Garland, Midford, & Ives, 1999; Schluter, Price, Moores, & Ludwig, 1997). Additionally, ancestral reconstruction may help contextualize observed patterns such as correlated shifts between phenotypic and environmental variables. Principles of ancestral state reconstruction may also be used to perform phylogenetic prediction, in which phenotypic values for unobserved or incompletely sampled taxa are estimated based on the evolutionary model and relative phylogenetic position (Garland & Ives, 2000).

Several methods have been developed to reconstruct ancestral phenotypes, including parsimony-based, Bayesian methods, and maximum likelihood (ML) estimation, the latter of which constitutes the focus of this paper (Felsenstein, 1985; Maddison, 1991; Revell & Reynolds, 2012; Schluter et al., 1997). Like other phylogenetic comparative methods, ancestral state reconstruction becomes increasingly time-consuming and computationally demanding as the number of species increases. Although efficient algorithms for most applications have existed since the initial development of modern comparative methods, their importance has recently seen a renewed emphasis (FitzJohn, 2012; Freckleton, 2012; Ho & Ané, 2014). Fast comparative methods are critical to keeping up with the ever-increasing size of phylogenetic trees in studies, as well as for statistical methods requiring thousands or millions of repeated calculations (e.g., parametric bootstrapping, Bayesian inference) (Boettiger & Ralph, 2012; Goolsby, 2016; Hadfield & Nakagawa, 2010; Schluter et al., 1997).

Unlike most comparative methods (e.g., phylogenetic regression, phylogenetic signal, estimation of alternative evolutionary models), computationally efficient methods for performing ancestral state reconstruction are severely lacking. This is because, despite the existence of efficient comparative methods that avoid the need to invert the phylogenetic covariance matrix, most R implementations of ML ancestral state reconstruction rely on (1) rerooting the tree at each internal node and performing repeated calculations (Revell, 2012), (2) high-dimensional numerical optimization (Paradis, Claude, & Strimmer, 2004), or (3) parameterizing and manipulating extremely large covariance matrices (Ho & Ané, 2014; Paradis et al., 2004).

This paper introduces a computationally efficient, generalizable, two-pass algorithm for performing ML ancestral state reconstruction which outperforms existing implementations by several orders of magnitude. The algorithm is first described in univariate terms and is mathematically identical to efficient algorithms described by Maddison (1991), Felsenstein (2004), and Elliot (2015). Next, the algorithm is generalized to multivariate trait evolution, non-Brownian models, missing data, and within-species variation.

The first pass of the algorithm is identical to the linear-time algorithm described in Ho and Ané (2014), which computes quantities at the root of the tree using a postorder (tips to root) tree traversal algorithm. The second pass of the algorithm operates by holding values computed at the root constant and recursively traversing the tree in preorder (root to tips) to compute quantities of interest at each internal node. The algorithm is implemented in the R package *Rphylopars* in the functions *anc.recon* and *phylopars* (Goolsby, Bruggeman, & Ané, 2017).

2 | METHODS

2.1 | Fast algorithm for ML ancestral state reconstruction

Here, we define a two-pass (postorder–preorder) recursive algorithm for calculating several quantities of interest related to ML ancestral state reconstruction at each node of the tree. The postorder portion of the algorithm as described in Ho and Ané (2014) partitions the phylogeny into recursively defined subtrees. For a terminal node (a tip) on the tree, the corresponding subtree consists of a single node (i.e., the tip of the subtree is also the root of the subtree), and the edge giving rise to the tip on the original phylogeny is the root edge of the subtree. For a bifurcating internal node, the corresponding subtree has two tips and a single internal node with a root edge (for a polytomous internal node, the subtree has multiple tips and a root edge). Like the PIC algorithm (Felsenstein, 1985), the postorder portion of the algorithm recursively computes locally parsimonious values for quantities of interest, including the expected variance due to phylogeny and estimated ancestral states at each internal node (Ho & Ané, 2014). In other words, local quantities that are calculated for a given node represent the global quantities that would be obtained if the tree consisted only of the given node and its descendants. At the root of the original phylogeny, the computed local quantity is equivalent to the global quantity, corresponding to globally parsimonious and maximum likelihood estimates (Felsenstein, 1985; Garland et al., 1999; Ho & Ané, 2014; Maddison, 1991). Conversely, the local quantities obtained for all other internal nodes are *not* global quantities because they do not account for the information contained in the rest of the phylogeny. However, because the postorder algorithm computes global quantities for the root of the tree, we can hold the root quantities constant and solve for values at its descendent nodes, which can then themselves be held constant to solve for their descendent nodes, and so on, until we reach the tips of the tree. The two-pass algorithm is mathematically equivalent to rerooting strategies for obtaining global estimates for each node (which are the current method-of-choice for rapid ancestral state reconstruction in R (Revell, 2012)), but the proposed algorithm avoids redundant time-consuming operations and is accordingly several orders of magnitude faster.

The two-pass algorithm described here computes the following quantities: $\hat{\boldsymbol{\mu}}^{(e)} = \left(\mathbf{1}'\mathbf{C}^{(e)^{-1}}\mathbf{1}\right)^{-1}\mathbf{1}'\mathbf{C}^{(e)^{-1}}\mathbf{Y}$, $p^{(e)} = \mathbf{1}'\mathbf{C}^{(e)^{-1}}\mathbf{1}$, $\mathbf{Q}^{(e)} = \mathbf{L}'\mathbf{C}^{(e)^{-1}}\mathbf{R}$, and the log determinant of the species covariance matrix ($\log|\mathbf{C}^{(e)}|$), where $\mathbf{1}$ is a vector of ones, $\hat{\boldsymbol{\mu}}^{(e)}$ is the ML ancestral estimate for $\mathbf{Y}$ at the node arising from edge e, $\mathbf{C}^{(e)}$ is the species covariance matrix obtained by rerooting the phylogeny at the node arising from edge ***e***, and **L** and **R** are matrices of compatible dimensions in the product $\mathbf{L}'\mathbf{C}^{(e)^{-1}}\mathbf{R}$ (e.g., $\mathbf{L}=\mathbf{1}$ and $\mathbf{R}=\mathbf{Y}$). These quantities are computed via preorder tree traversal following postorder computation of the local quantities $\tilde{\boldsymbol{\mu}}^{(e)} = \left(\mathbf{1}'\tilde{\mathbf{C}}^{(e)^{-1}}\mathbf{1}\right)^{-1}\mathbf{1}'\tilde{\mathbf{C}}^{(e)^{-1}}\mathbf{Y}$, $\tilde{p}^{(e)} = \mathbf{1}'\tilde{\mathbf{C}}^{(e)^{-1}}\mathbf{1}$, $\tilde{\mathbf{Q}}^{(e)} = \mathbf{L}'\tilde{\mathbf{C}}^{(e)^{-1}}\mathbf{R}$, and $\log|\tilde{\mathbf{C}}^{(e)}|$, where $\tilde{\mathbf{C}}^{(e)}$ is the species covariance matrix obtained by pruning the tree to only the descendants arising from (but not including) edge ***e***. Note that $1/\tilde{p}^{(e)}$ is equivalent to the transformed branch lengths obtained using the phylogenetically independent contrasts (PIC) algorithm and $\tilde{\boldsymbol{\mu}}^{(e)}$ is equivalent to PIC-based (locally parsimonious) ancestral state reconstruction (Felsenstein, 1985).

1. Initialization: for edge e of length $t^{(e)}$ giving rise to a terminal taxon, define as follows:

$$\tilde{\boldsymbol{\mu}}^{(e)} = \mathbf{Y}^{(e)}$$

$$\tilde{p}^{(e)} = 1/t^{(e)}$$

$$\tilde{\mathbf{U}}^{(e)'} = \mathbf{L}^{(e)'}/t^{(e)}$$

$$\tilde{\mathbf{V}}^{(e)} = \mathbf{R}^{(e)}/t^{(e)}$$

$$\tilde{\mathbf{Q}}_e = \mathbf{L}^{(e)'}\mathbf{R}^{(e)}/t^{(e)}$$

$$\log|\tilde{\mathbf{C}}^{(e)}| = \log\left(t^{(e)}\right)$$

2. Postorder recursion: for edge e of length $t^{(e)}$ giving rise to an internal node, define for all immediate descendants (d) of edge e:

$$p_A^{(e)} = \Sigma\tilde{p}^{(d)}$$

$$\tilde{\boldsymbol{\mu}}^{(e)} = \Sigma\left(\tilde{\boldsymbol{\mu}}^{(d)}\tilde{p}^{(d)}/p_A^{(e)}\right)$$

$$\tilde{p}^{(e)} = p_A^{(e)}/\left(1+t^{(e)}p_A^{(e)}\right)$$

$$\tilde{\mathbf{U}}^{(e)'} = \left(\Sigma\tilde{\mathbf{U}}^{(d)'}\right)/\left(1+t^{(e)}p_A^{(e)}\right)$$

$$\tilde{\mathbf{V}}^{(e)} = \left(\Sigma\tilde{\mathbf{V}}^{(d)}\right)/\left(1+t^{(e)}p_A^{(e)}\right)$$

$$\tilde{\mathbf{Q}}^{(e)} = \Sigma\tilde{\mathbf{Q}}^{(d)} - \left(\Sigma\tilde{\mathbf{U}}^{(d)'}\right)\left(\Sigma\tilde{\mathbf{V}}^{(d)}\right)t^{(e)}/\left(1+t^{(e)}p_A^{(e)}\right)$$

$$\log|\tilde{\mathbf{C}}_e| = \Sigma\log|\tilde{\mathbf{C}}^{(d)}| + \log\left(1+t^{(e)}p_A^{(e)}\right)$$

3. At the root edge of the tree, denote as follows:

$$\hat{\boldsymbol{\mu}}^{(r)} = \tilde{\boldsymbol{\mu}}^{(r)}$$

$$p^{(r)} = \tilde{p}^{(r)}$$

$$\mathbf{U}^{(r)'} = \tilde{\mathbf{U}}^{(r)'}$$

$$\mathbf{V}^{(r)} = \tilde{\mathbf{V}}^{(r)}$$

$$\mathbf{Q}^{(r)} = \tilde{\mathbf{Q}}^{(r)}$$

$$\log|\mathbf{C}^{(r)}| = \log|\tilde{\mathbf{C}}^{(r)}|$$

4. Preorder recursion: for edge e (which arises from the node arising from ancestral edge a) of length t_e giving rise to an internal node, define as follows:

$$\hat{\boldsymbol{\mu}}^{(e)} = \tilde{\boldsymbol{\mu}}^{(e)}\tilde{p}^{(e)}t^{(e)} + \hat{\boldsymbol{\mu}}^{(a)} - \hat{\boldsymbol{\mu}}^{(a)}\tilde{p}^{(e)}t^{(e)}$$

$$p^{(e)} = \tilde{p}^{(e)}/\left(1-t^{(e)}\tilde{p}^{(e)}\right) + \left(p^{(a)}-\tilde{p}^{(e)}\right)/\left(1+t^{(e)}\left(p^{(a)}-\tilde{p}^{(e)}\right)\right)$$

$$\mathbf{U}^{(e)'} = \tilde{\mathbf{U}}^{(e)'}/\left(1-t^{(e)}\tilde{p}^{(e)}\right) + \left(\mathbf{U}^{(a)'}-\tilde{\mathbf{U}}^{(e)'}\right)/\left(1+t^{(e)}\left(p^{(a)}-\tilde{p}^{(e)}\right)\right)$$

$$\mathbf{V}^{(e)} = \tilde{\mathbf{V}}^{(e)}/\left(1-t^{(e)}\tilde{p}^{(e)}\right) + \left(\mathbf{V}^{(a)}-\tilde{\mathbf{V}}^{(e)}\right)/\left(1+t^{(e)}\left(p^{(a)}-\tilde{p}^{(e)}\right)\right)$$

$$\mathbf{Q}^{(e)} = \left[\tilde{\mathbf{Q}}^{(e)} - \left(\tilde{\mathbf{U}}^{(e)'}\right)\left(\tilde{\mathbf{V}}^{(e)}\right)\left(-t^{(e)}\right)/\left(1-t^{(e)}\tilde{p}^{(e)}\right)\right] + \left[\left(\mathbf{Q}^{(a)}-\tilde{\mathbf{Q}}^{(e)}\right) - \left(\mathbf{U}^{(a)'}-\tilde{\mathbf{U}}^{(e)'}\right)\left(\mathbf{V}^{(a)}-\tilde{\mathbf{V}}^{(e)}\right)t^{(e)}/\left(1+t^{(e)}\left(p^{(a)}-\tilde{p}^{(e)}\right)\right)\right]$$

$$\log|\mathbf{C}^{(e)}| = \log|\mathbf{C}^{(a)}| + \log\left(1-t^{(e)}\tilde{p}^{(e)}\right) + \log\left(1+t^{(e)}\left(p^{(a)}-\tilde{p}^{(e)}\right)\right)$$

For a linear regression model, we may also compute the regression parameters $\hat{\boldsymbol{\beta}}^{(e)} = \left(\mathbf{Q}_{\mathbf{XX}}^{(e)}\right)^{-1}\mathbf{Q}_{\mathbf{XY}}^{(e)}$ where **X** is a design matrix (for an intercept-only model, **X** = **1** as above; for a regression model, the first column typically consists of ones and the remaining columns consist of values for predictor variables).

Ho and Ané (2014) proved that the postorder recursion algorithm yields the global quantities $\hat{\boldsymbol{\mu}}^{(r)}, p^{(r)}, \mathbf{Q}^{(r)}$, and $\log|\mathbf{C}^{(r)}|$, and it has been long-established that rerooting the tree yields global estimates of these quantities for any node of the tree (Garland & Ives, 2000; Maddison, 1991; Swofford & Maddison, 1987). The preorder recursion step works by mathematically rerooting each subtree recursively at each node. To demonstrate the properties of the preorder recursion, first consider that the original phylogeny lacks a root edge ($t^{(r)} = 0$), so step 3 reduces to $p^{(r)} = \Sigma\tilde{p}^{(d)}$. Therefore, we may treat the current subtree as being composed of two descendent edges which we denote e and *other*, such that $p^{(r)} = \Sigma\tilde{p}^{(d)} = \tilde{p}^{(e)} + \tilde{p}^{(\text{other})}$, which can also be expressed as $p^{(r)} = \tilde{p}^{(e)} + \left(p^{(r)} - \tilde{p}^{(e)}\right)$ to avoid having to keep track of $\tilde{p}^{(\text{other})}$ (note that this holds true even if the subtree arising from *other* were in fact polytomous).

To compute the quantity $p^{(e)}$, we could reroot the original tree at the node arising from edge e and then perform steps 1–3 of the postorder algorithm (Garland et al., 1999; Ho & Ané, 2014). However, the majority of these steps would be redundant, as we have already computed all of these quantities up to our node of interest. To see this, note that had the original tree been rooted at the node arising from edge e rather than r, the original computation for $\tilde{p}^{(e)}$ would have been $\tilde{p}^{(e)^*} = p_A^{(e)}$ instead of $\tilde{p}^{(e)} = p_A^{(e)}/\left(1+t^{(e)}p_A^{(e)}\right)$ because $t^{(e)}$ would have equaled zero (the length of $t^{(e)}$ would have instead been added to the length of $t^{(\text{other})}$). To adjust for this, we cancel out the contribution of $t^{(e)}$ as follows: $\tilde{p}^{(e)^*} = \tilde{p}^{(e)}/\left(1-t^{(e)}\tilde{p}^{(e)}\right)$. Now, we must add the contribution of $t^{(e)}$ to $\tilde{p}^{(\text{other})}$, as follows: $\tilde{p}^{(\text{other})^*} = \tilde{p}^{(\text{other})}/\left(1+t^{(e)}\tilde{p}^{(\text{other})}\right) = \left(p^{(r)}-\tilde{p}^{(e)}\right)/\left(1+t^{(e)}\left(p^{(r)}-\tilde{p}^{(e)}\right)\right)$. Therefore, we have now obtained the quantities necessary to compute $p^{(r)^*}$ (i.e., had the tree been rooted at the node arising from edge e) without actually having to reroot the tree or perform any redundant calculations: $p^{(e)} = \tilde{p}^{(e)^*} + \tilde{p}^{(\text{other})^*} = \tilde{p}^{(e)}/\left(1-t^{(e)}\tilde{p}^{(e)}\right) + \left(p^{(a)}-\tilde{p}^{(e)}\right)/\left(1+t^{(e)}\left(p^{(a)}-\tilde{p}^{(e)}\right)\right)$.

The same procedure immediately follows for the computations of $\mathbf{U}^{(e)\prime}$, $V^{(e)}$, $Q^{(e)}$, and $\log|C^{(e)}|$ because at the root of the tree, these quantities are simply composed of the sums of their descendant quantities (because $t^{(r)} = 0$). The ancestral state reconstruction $\hat{\boldsymbol{\mu}}^{(e)}$ is an algebraic simplification of the quantity $\hat{\boldsymbol{\mu}}^{(e)} = \left(\mathbf{Q}_{11}^{(e)}\right)^{-1}\mathbf{Q}_{1Y}^{(e)}$ where $\mathbf{Q}_{11}^{(e)} = \mathbf{1}'\mathbf{C}^{(e)^{-1}}\mathbf{1}$ and $\mathbf{Q}_{1Y}^{(e)} = \mathbf{1}'\mathbf{C}^{(e)^{-1}}\mathbf{Y}$. By repeating step 4 recursively from the root to the tips, we obtain global ML estimates for each internal node.

The covariance of a given estimate can be computed as follows: $\mathrm{cov}_{\hat{\boldsymbol{\mu}}^{(e)}} = \hat{\boldsymbol{\Sigma}}/p^{(e)}$ where $\hat{\boldsymbol{\Sigma}}$ is the ML or restricted ML evolutionary rate matrix: $\hat{\boldsymbol{\Sigma}} = \left(\left(\mathbf{Y} - \mathbf{1}\hat{\boldsymbol{\mu}}^{(r)}\right)'\mathbf{C}^{(r)^{-1}}\left(\mathbf{Y} - \mathbf{1}\hat{\boldsymbol{\mu}}^{(r)}\right)\right)/(N - \mathrm{REML})$, N is the number of species, and REML = 1 if the restricted ML estimate is desired and REML = 0 otherwise. 95% confidence intervals for an estimate can then be computed as follows: $95\%\,\mathrm{C.I.}_{\hat{\boldsymbol{\mu}}^{(e)}} = \hat{\boldsymbol{\mu}}^{(e)} \pm 1.96\sqrt{\mathrm{var}_{\hat{\boldsymbol{\mu}}^{(e)}}}$ (Garland & Ives, 2000; Garland et al., 1999; Rohlf, 2001).

2.2 | Multivariate data, alternative evolutionary models, within-species variation, and missing data

The described algorithm can be easily modified to incorporate a wide variety of models with various features such as missing data, intraspecific variation, and alternative evolutionary model specifications (Bruggeman, Heringa, & Brandt, 2009; Felsenstein, 2008; Goolsby et al., 2017; Ives, Midford, & Garland, 2007). For a multivariate model of evolution, the N × M matrix $\mathbf{Y}$ (where M is the number of traits) is rearranged into an NM-length column vector y, the matrix $\mathbf{1}$ is replaced with an NM × M matrix describing which observations of $\mathbf{y}$ are from which trait, and the covariance of each observation is described by an NM × NM species-trait covariance matrix $\mathbf{W}$. For a Brownian motion model of evolution, $\mathbf{W} = \boldsymbol{\Sigma} \otimes \mathbf{C}$, where $\boldsymbol{\Sigma}$ is the evolutionary rate matrix, ⊗ denotes the Kronecker product, and W is partitioned into M^2 blocks of size N × N. For example, at block i, j, $\mathbf{W}_{i,j} = \Sigma_{i,}^{\,j}\mathbf{C}$. When considering the node arising from a single edge e, we are left with an M × M matrix of transformed heights (root-to-node distance): $\mathbf{H}^{(e)} = C_{a,b}\boldsymbol{\Sigma}$, and the node arising form edge e is the most recent common ancestor of species a and b. The height matrix $\mathbf{H}_{\mathbf{e}}$ can be converted into an edge length matrix $\mathbf{T}^{(e)}$ as follows: $\mathbf{T}^{(e)} = \mathbf{H}^{(a)} - \mathbf{H}^{(e)}$ (which also equals $t^{(e)}\boldsymbol{\Sigma}$ for a Brownian motion model), where the node arising from edge a is the parent of edge e. For Brownian motion models, we can simply use $\mathbf{T}^{(e)} = t^{(e)}\boldsymbol{\Sigma}$. To accommodate rate shift models, the estimated regime-specific rate matrices $\boldsymbol{\Sigma}^{(s)}$ may be used: $\mathbf{T}^{(e)} = t^{(e)}\boldsymbol{\Sigma}^{(s)}$. For more complex evolutionary models (e.g., multivariate Ornstein–Uhlenbeck on an ultrametric tree), $\mathbf{W}$ is scaled according to block-specific transformations, and we must compute $\mathbf{T}^{(e)} = \mathbf{H}^{(a)} - \mathbf{H}^{(e)}$ for each edge (for a derivation, see Goolsby et al., 2017; Appendix S1). It should be noted that the algorithm requires an ultrametric tree if an Ornstein–Uhlenbeck model is specified; otherwise, a complex series of branch length and data transformations must be made to maintain three-point structure as described in Ho and Ané (2014). The multivariate algorithm proceeds as follows:

1. Initialization: for edge e with length matrix $\mathbf{T}^{(e)}$ giving rise to a terminal taxon, for the subset of variables $\mathbf{k}$ on which data are available (nonmissing)

$$\tilde{\mathbf{p}}_{\mathbf{k,k}}^{(e)} = \mathbf{T}_{\mathbf{k,k}}^{(e)^{-1}}$$

$\tilde{\mathbf{U}}_{\mathbf{k}}^{(e)\prime} = \mathbf{L}_{\mathbf{k}}^{(e)\prime}\tilde{\mathbf{p}}_{\mathbf{k,k}}^{(e)}$ for the subset of variables on which data are available. Rows of $\tilde{\mathbf{U}}^{(e)}$ corresponding to missing data are set to zero.

$\tilde{\mathbf{V}}_{\mathbf{k}}^{(e)} = \tilde{\mathbf{p}}_{\mathbf{k,k}}^{(e)}\mathbf{R}_{\mathbf{k}}^{(e)}$ for the subset of variables on which data are available. Columns of $\tilde{\mathbf{V}}^{(e)}$ corresponding to variables with missing data are set to zero.

$\tilde{\mathbf{Q}}_{\mathbf{k,k}}^{(e)} = \mathbf{L}_{\mathbf{k}}^{(e)\prime}\tilde{\mathbf{p}}_{\mathbf{k,k}}^{(e)}\mathbf{R}_{\mathbf{k}}^{(e)}$ for the subset of variables on which data are available. Rows and columns of $\tilde{\mathbf{Q}}^{(e)}$ corresponding to missing data are set to zero.

$\log|\tilde{\mathbf{W}}^{(e)}| = \log|\mathbf{T}_{\mathbf{k,k}}^{(e)}|$ for the subset of variables on which data are available.

2. Postorder recursion: for edge e with length matrix $\mathbf{T}^{(e)}$ giving rise to an internal node, define for all immediate descendants (d) of edge e:

$$\mathbf{p}_A^{(e)} = \Sigma\tilde{\mathbf{p}}^{(d)}$$

$$\tilde{\mathbf{p}}^{(e)} = \mathbf{p}_A^{(e)}\left(\mathbf{I} + \mathbf{T}^{(e)}\mathbf{p}_A^{(e)}\right)^{-1}$$

$$\tilde{\mathbf{U}}^{(e)\prime} = \left(\Sigma\tilde{\mathbf{U}}^{(d)\prime}\right)\left(\mathbf{I} + \mathbf{T}^{(e)}\mathbf{p}_A^{(e)}\right)^{-1}$$

$$\tilde{\mathbf{V}}^{(e)} = \left(\left(\Sigma\tilde{\mathbf{V}}^{(d)}\right)'\left(\mathbf{I} + \mathbf{T}^{(e)}\mathbf{p}_A^{(e)}\right)^{-1}\right)'$$

$$\tilde{\mathbf{Q}}^{(e)} = \Sigma\tilde{\mathbf{Q}}^{(d)} - \left(\Sigma\tilde{\mathbf{U}}^{(d)}\right)\left(\mathbf{I} + \mathbf{T}^{(e)}\mathbf{p}_A^{(e)}\right)^{-1}\mathbf{T}^{(e)}(\Sigma\tilde{\mathbf{V}}^{(d)})$$

$$\log|\tilde{\mathbf{W}}^{(e)}| = \Sigma\log|\tilde{\mathbf{W}}^{(d)}| + \log|\mathbf{I} + \mathbf{T}^{(e)}\mathbf{p}_A^{(e)}|$$

3. At the root edge of the tree, denote:

$$\mathbf{p}^{(r)} = \tilde{\mathbf{p}}^{(r)}$$

$$\mathbf{U}^{(r)\prime} = \tilde{\mathbf{U}}^{(r)\prime}$$

$$\mathbf{V}^{(r)} = \tilde{\mathbf{V}}^{(r)}$$

$$\mathbf{Q}^{(r)} = \tilde{\mathbf{Q}}^{(r)}$$

$$\log|\mathbf{W}^{(r)}| = \log|\tilde{\mathbf{W}}^{(e)}|$$

$$\hat{\boldsymbol{\mu}}^{(r)} = \left(\mathbf{Q}_{11}^{(r)}\right)^{-1}\mathbf{Q}_{1Y}^{(r)}$$

4. Preorder recursion: for edge e (which arises from the node arising from ancestral edge a) of length $\mathbf{T}^{(e)}$ giving rise to an internal node, define

$$\mathbf{p}^{(e)} = \tilde{\mathbf{p}}^{(e)}\left(\mathbf{I} - \mathbf{T}^{(e)}\tilde{\mathbf{p}}^{(e)}\right)^{-1} + \left(\mathbf{p}^{(a)} - \tilde{\mathbf{p}}^{(e)}\right)\left(\mathbf{I} + \mathbf{T}^{(e)}\left(\mathbf{p}^{(a)} - \tilde{\mathbf{p}}^{(e)}\right)\right)^{-1}$$

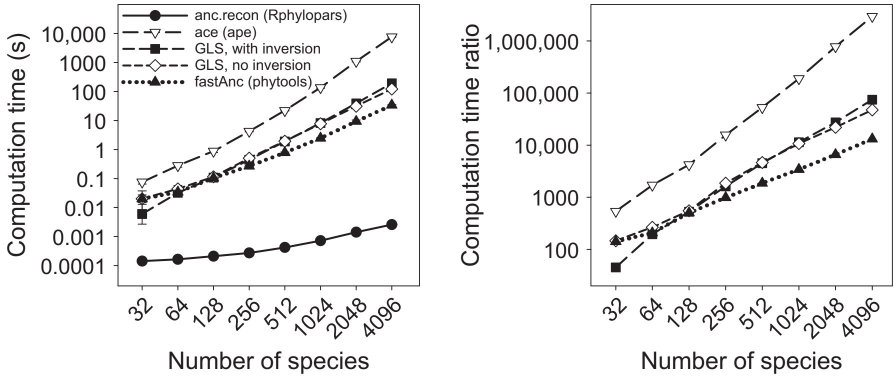

FIGURE 1 Computation times (left) for univariate ancestral state reconstruction using the described fast two-pass algorithm (*anc.recon* function, *Rphylopars* package), numerical optimization (*ace* function, *ape* package), generalized least squares (GLS) with matrix inversion (Martins & Hansen, 1997), GLS without matrix inversion (Ho & Ané, 2014), and the rerooting method implemented in the *fastAnc* function in *phytools*. The right panel consists of ratios of computation times for optimization, GLS with and without inversion, and rerooting relative to the described fast algorithm. All *anc.recon* run times completed in fewer than 10 ms, whereas the next-fastest method (*fastAnc*) ran from 141 to 13,104 times slower than *anc.recon*, and the slowest method (*ace*) ranged from 537 to nearly three million times slower than *anc.recon* (right panel). Error bars (where visible) indicate standard deviation of five replicate runs per simulated number of species

$$\mathbf{U}^{(e)\prime} = \tilde{\mathbf{U}}^{(e)\prime}\left(\mathbf{I} - \mathbf{T}^{(e)}\tilde{\mathbf{p}}^{(e)}\right)^{-1} + \left(\mathbf{U}^{(a)\prime} - \tilde{\mathbf{U}}^{(e)\prime}\right)\left(\mathbf{I} + \mathbf{T}^{(e)}\left(\mathbf{p}^{(a)} - \tilde{\mathbf{p}}^{(e)}\right)\right)^{-1}$$

$$\mathbf{V}^{(e)} = \left[\tilde{\mathbf{V}}^{(e)\prime}\left(\mathbf{I} - \mathbf{T}^{(e)}\tilde{\mathbf{p}}^{(e)}\right)^{-1} + \left(\mathbf{V}^{(a)} - \tilde{\mathbf{V}}^{(e)}\right)^{\prime}\left(\mathbf{I} + \mathbf{T}^{(e)}\left(\mathbf{p}^{(a)} - \tilde{\mathbf{p}}^{(e)}\right)\right)^{-1}\right]^{\prime}$$

$$\mathbf{Q}^{(e)} = \left[\tilde{\mathbf{Q}}^{(e)} - \left(\tilde{\mathbf{U}}^{(e)\prime}\right)\left(\tilde{\mathbf{V}}^{(e)}\right)\left(-\mathbf{T}^{(e)}\right)\right]\left(\mathbf{I} - \mathbf{T}^{(e)}\tilde{\mathbf{p}}^{(e)}\right)^{-1} + \left[\left(\mathbf{Q}^{(a)} - \tilde{\mathbf{Q}}^{(e)}\right) - \left(\mathbf{U}^{(a)\prime} - \tilde{\mathbf{U}}^{(e)\prime}\right)\left(\mathbf{I} + \mathbf{T}^{(e)}\left(\mathbf{p}^{(a)} - \tilde{\mathbf{p}}^{(e)}\right)\right)^{-1}\mathbf{T}^{(e)}\left(\mathbf{V}^{(a)} - \tilde{\mathbf{V}}^{(e)}\right)\right]$$

$$\log|\mathbf{W}_e| = \log|\mathbf{W}_a| + \log|\mathbf{I} - \mathbf{T}^{(e)}\tilde{\mathbf{p}}^{(e)}| + \log|\mathbf{I} + \mathbf{T}^{(e)}(\mathbf{p}^{(a)} - \tilde{\mathbf{p}}^{(e)})|$$

$$\hat{\boldsymbol{\mu}}^{(e)} = \left(\mathbf{Q}_{11}^{(e)}\right)^{-1}\mathbf{Q}_{1Y}^{(e)}$$

5. For edge e (which arises from the node arising from ancestral edge a) of length $\mathbf{T}^{(e)}$ giving rise to a terminal node (a tip) with missing data on a subset of variables ($\mathbf{u}$) and nonmissing data for subset $\mathbf{k}$, define

$$\hat{\boldsymbol{\mu}}_{\mathbf{u}}^{(e)} = \mathbf{T}_{\mathbf{u},\mathbf{k}}^{(e)}\mathbf{T}_{\mathbf{k},\mathbf{k}}^{(e)-1}\left(\mathbf{Y}_{\mathbf{k}}^{(e)} - \hat{\boldsymbol{\mu}}_{\mathbf{k}}^{(a)}\right)^{\prime} + \hat{\boldsymbol{\mu}}_{\mathbf{u}}^{(a)}$$

$$\mathrm{cov}_{\hat{\boldsymbol{\mu}}_{\mathbf{u}}^{(e)}} = \left[\left(\mathbf{p}^{(a)} - \tilde{\mathbf{p}}^{(e)}\right)\left(\mathbf{I} + \mathbf{T}^{(e)}\tilde{\mathbf{p}}^{(e)}\right)^{-1}\right]_{\mathbf{u},\mathbf{u}}^{-1}$$

To accommodate within-species variation when only summary data are available, the above algorithm is identical except that in steps 1, $\mathbf{T}^{(e)}$ is replaced with $\mathbf{T}^{(e)} + \mathbf{B}^{(e)}$ where $\mathbf{B}^{(e)}$ is an estimate of within-species covariance (e.g., a diagonal matrix with squared standard errors) (Ives et al., 2007). For species mean imputation in step 5, $\mathbf{B}^{(e)}$ is *not* added to $\mathbf{T}^{(e)}$ (Bruggeman et al., 2009; Goolsby et al., 2017).

To accommodate within-species variation when raw data are available, the algorithm is nearly identical as above except that initialization (step 1) and imputation of missing data (step 5) is performed on raw data (i.e., an individual within-species observation) rather than on species means, and $\mathbf{T}^{(e)}$ replaced entirely with $\mathbf{B}^{(e)}$ in steps 1 and 5. $\mathbf{B}^{(e)}$ may be set to an a priori determined value (Ives et al., 2007) or jointly estimated during maximum likelihood optimization (Felsenstein, 2008). Typically, $\mathbf{B}^{(e)}$ is assumed to be identical across species if $\mathbf{B}^{(e)}$ is to be estimated via numerical optimization (Felsenstein, 2008). Steps 2–4 proceed as normal, except that species nodes are treated as "internal nodes" since the "tips" of the tree are individual observations, and hence edges giving rise to species nodes are included in the postorder and preorder recursion steps. When e gives rise to a species node, step 4 provides estimates of species means, and step 5 provides raw data imputations for missing values.

TABLE 1 Mean computation times for *anc.recon* ancestral state reconstruction on univariate datasets with 256 to 2,097,152 (2^8 to 2^{21}) species. For each number of species, five simulated phylogenies and datasets were generated

Number of species	Computation time (s)	Standard deviation
256	0.0003	1.87E-05
512	0.0004	1.67E-05
1,024	0.0007	1.14E-05
2,048	0.001	3.36E-05
4,096	0.003	8.29E-05
8,192	0.006	0.0004
16,384	0.011	0.0006
32,768	0.021	0.0004
65,536	0.052	0.0084
131,072	0.110	0.0071
262,144	0.222	0.0148
524,288	0.520	0.0418
1,048,576	1.136	0.0929
2,097,152	2.422	0.4268

3 | RESULTS AND DISCUSSION

3.1 | R implementation

The proposed ancestral state reconstruction algorithm is implemented in the R package *Rphylopars* (Goolsby et al., 2017). For simple Brownian motion evolution on univariate or multivariate data, maximum likelihood ancestral states and confidence intervals may be fit using the *Rphylopars* function *anc.recon*. For more complex models with missing data, within-species variation, or alternative evolutionary model specifications (e.g., Ornstein–Uhlenbeck or Early-Burst), the *Rphylopars* function *phylopars* must be used to fit evolutionary model parameters, which are then used to compute maximum likelihood ancestral states using the fast algorithm.

3.2 | Speed comparisons: univariate data

Here, we compare the speed of the proposed algorithm is implemented in *anc.recon* with four standard methods as implemented in R for performing ML ancestral state reconstruction: (1) numerical optimization (*ace* function in the R package *ape*, Paradis et al., 2004), (2) generalized least squares with direct matrix inversion (Martins & Hansen, 1997), (3) generalized least squares avoiding matrix inversion using the linear-time algorithm described in Ho and Ané (2014), and (4) the rerooting method implemented in the *fastAnc* function in the *phytools* package (Revell, 2012). Univariate traits were simulated on phylogenies of size 32, 64, 128, 256, 512, 1,024, 2,048, and 4,096 species using the *rTraitCont* and *rtree* functions in *ape* (Paradis et al., 2004). For each tree size, five simulated phylogenies and datasets were generated, and the mean and standard deviation of computation time was recorded for each method. In order to be able to distinguish the computation time of the algorithm described here from 0 s (using the *system.time* function, which has a resolution of 10 ms), speed assessments using *anc.recon* were performed on 1,000 replicated function calls and the total computation time was subsequently divided by 1,000.

For all simulated datasets, *anc.recon* computation time was below 10 ms, whereas the *fastanc* function took up to 36 s for the largest simulated dataset (4,096 taxa), with a polynomial increase in computation time as the number of species increased (Figure 1a). Other methods were even slower, including numerical optimization, in which *anc.recon* performed approximately 3,000,000 times faster than *ace* (Figure 1b). Even on the smallest simulated datasets (32 taxa), *anc.recon* was approximately 140 times faster than *fastAnc* (the next fastest method), and for the largest dataset, *anc.recon* was over 13,000 times faster than *fastAnc*. Additionally, a decrease in precision was observed for numerical optimization in the *ace* function, something not shared by the method described here (which algorithmically computes exact maximum likelihood estimates). Speed assessments were also performed using only *anc.recon* on phylogenies ranging from 256 to 2,097,152 (2^8 to 2^{21}) taxa, the largest of which completed in fewer than 3 s. Across all simulations, *anc.recon* exhibited linear increases in computation time (Table 1). R code for performing the simulations used to generate all figures is supplied in Appendix S1.

4 | CONCLUSION

The algorithm described here generalizes existing efficient algorithms (Elliot, 2015; Felsenstein, 2004; Maddison, 1991) and is capable of performing maximum likelihood ancestral state reconstruction on phylogenies containing one million taxa in fewer than 2 s, using modest computational resources (i.e., a standard laptop). The method can be expanded to incorporate a variety of models, including multivariate generalizations, within-species variation, non-Brownian evolutionary models, rate heterogeneity, and more. As the number of taxa in phylogenetic comparative studies continues to rise, efficient linear-time algorithms will become increasingly critical. Additionally, frameworks requiring thousands or millions of repeated calculations, such as parametric bootstrapping and Bayesian analyses, will also benefit from the continued improvement of fast algorithms.

ACKNOWLEDGMENTS

EWG wishes to thank Cécile Ané and Devon P. Humphreys for thought-provoking discussions on efficient phylogenetic comparative methods, as well two anonymous reviewers for helpful feedback. This work was supported in part by the National Science Foundation [DEB-1501215].

CONFLICT OF INTEREST

None declared.

REFERENCES

Boettiger, C., & Ralph, P. (2012). Is your phylogeny informative? Measuring the power of comparative methods. *Evolution*, *66*, 2240–2251.

Bruggeman, J., Heringa, J., & Brandt, B. W. (2009). PhyloPars: Estimation of missing parameter values using phylogeny. *Nucleic Acids Research*, *37*, W179–W18.

Elliot, M. J. (2015). Identical inferences about correlated evolution arise from ancestral state reconstruction and independent contrasts. *Journal of Theoretical Biology*, *364*, 321–325.

Felsenstein, J. (1985). Phylogenies and the comparative method. *American Naturalist*, *125*, 1–15.

Felsenstein, J. (2004). Brownian motion and gene frequencies. In *Inferring phylogenies* (pp. 391–414). Sunderland, MA: Sinauer Associates.

Felsenstein, J. (2008). Comparative methods with sampling error and within-species variation: Contrasts revisited and revised. *American Naturalist*, *171*, 713–725.

FitzJohn, R. J. (2012). Diversitree: Comparative phylogenetic analyses of diversification in R. *Methods in Ecology & Evolution*, *3*, 1084–1092.

Freckleton, R. P. (2012). Fast likelihood calculations for comparative analyses. *Methods in Ecology & Evolution*, *3*, 940–947.

Garland, T., & Ives, A. R. (2000). Using the past to predict the present: Confidence intervals for regression equations in phylogenetic comparative methods. *American Naturalist*, *155*, 346–364.

Garland, T., Midford, P. E., & Ives, A. R. (1999). An introduction to phylogenetically based statistical methods, with a new method for confidence intervals on ancestral values. *American Zoologist*, *39*, 374–388.

Goolsby, E. W. (2016). Likelihood-based parameter estimation for high-dimensional phylogenetic comparative models: Overcoming the limitations of 'distance-based' methods. *Systematic Biology*, *65*, 852–870.

Goolsby, E. W., Bruggeman, J., & Ané, C. (2017). Rphylopars: Fast multivariate phylogenetic comparative methods for missing data and within-species variation. *Methods in Ecology and Evolution*, *8*, 22–27.

Hadfield, J. D., & Nakagawa, S. (2010). General quantitative genetic methods for comparative biology: Phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *Journal of Evolutionary Biology*, *23*, 494–508.

Ho, L. T. H., & Ané, C. (2014). A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Systematic Biology*, *63*, 397–408.

Ives, A. R., Midford, P. E., & Garland, T. Jr (2007). Within-species variation and measurement error in phylogenetic comparative methods. *Systematic Biology*, *56*, 252–270.

Maddison, W. P. (1991). Squared-change parsimony reconstructions of ancestral states for continuous-valued characters on a phylogenetic tree. *Systematic Zoology*, *40*, 304–314.

Martins, E. P., & Hansen, T. F. (1997). Phylogenies and the comparative method: A general approach to incorporating phylogenetic information into the analysis of interspecific data. *American Naturalist*, *149*, 646–667.

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, *20*, 289–290.

Revell, L. J. (2012). phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology & Evolution*, *3*, 217–223.

Revell, L. J., & Reynolds, G. (2012). A new Bayesian method for fitting evolutionary models to comparative data with intraspecific variation. *Evolution*, *66*, 2697–2707.

Rohlf, F. J. (2001). Comparative methods for the analysis of continuous variables: Geometric interpretations. *Evolution*, *55*, 2143–2160.

Schluter, D., Price, T., Moores, A. O., & Ludwig, D. (1997). Likelihood of ancestral states in adaptive radiation. *Evolution*, *51*, 1699–1711.

Swofford, D. L., & Maddison, W. P. (1987). Reconstructing ancestral character states under Wagner parsimony. *Mathematical Biosciences*, *87*, 199–229.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Goolsby EW. Rapid maximum likelihood ancestral state reconstruction of continuous characters: A rerooting-free algorithm. *Ecol Evol*. 2017;00:1–7. https://doi.org/10.1002/ece3.2837