Accounting for ambiguity in ancestral sequence reconstruction

Adrien Oliva, Sylvain Pulicani, Vincent Lefort, Laurent Brehelin, Olivier Gascuel, Stéphane Guindon

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year

OXFORD

Phylogenetics

Accounting for ambiguity in ancestral sequence reconstruction

A. Oliva [1,2], S. Pulicani [1,], V. Lefort [1,], L. Bréhélin [1,], O. Gascuel [3], S. Guindon [1,*]

[1] LIRMM, CNRS & Université de Montpellier, Montpellier, France
[2] Australian Centre for Ancient DNA, Adelaide, Australia
[3] Unité Bioinformatique Evolutive, C3BI USR 3756, Institut Pasteur & CNRS, Paris, France

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: The reconstruction of ancestral genetic sequences from the analysis of contemporaneous data is a powerful tool to improve our understanding of molecular evolution. Various statistical criteria defined in a phylogenetic framework can be used to infer nucleotide, amino-acid or codon states at internal nodes of the tree, for every position along the sequence. These criteria generally select the state that maximizes (or minimizes) a given criterion. Although it is perfectly sensible from a statistical perspective, that strategy fails to convey useful information about the level of uncertainty associated to the inference.

Results: The present study introduces a new criterion for ancestral sequence reconstruction, the minimum posterior expected error (MPEE), that selects a single state whenever the signal conveyed by the data is strong, and a combination of multiple states otherwise. We also assess the performance of a criterion based on the Brier scoring scheme which, like MPEE, does not rely on any tuning parameters. The precision and accuracy of several other criteria that involve arbitrarily set tuning parameters are also evaluated. Large scale simulations demonstrate the benefits of using the MPEE and Brier-based criteria with a substantial increase in the accuracy of the inference of past sequences compared to the standard approach and realistic compromises on the precision of the solutions returned.

Availability: The software package PhyML (`https://github.com/stephaneguindon/phyml`) provides an implementation of the Maximum A Posteriori (MAP) and MPEE criteria for reconstructing ancestral nucleotide and amino-acid sequences.

Contact: `guindon@lirmm.fr`

1 Introduction

Molecular sequences collected in present day species provide a wealth of information about past evolutionary events. Using relevant probabilistic models of molecular evolution, it is possible to reconstruct the sequences of species ancestral to a sample of taxa (see Merkl and Sterner (2016) for a recent review). The application of these techniques led to spectacular findings. In particular, the resurrection of ancestral proteins using biochemical processes (Chang and Donoghue, 2000; Thornton, 2004; Bridgham et al., 2006) helped improved our understanding of the ways evolution takes place at the molecular level.

Phylogenetics provides an adequate framework for the reconstruction of ancestral sequences. Given a phylogenetic tree that depicts the evolutionary history of a sample of taxa along with a set of corresponding homologous sequences, it is possible to estimate the sequences at each internal node of the tree. The parsimony approach (Fitch, 1971) consists in selecting the ancestral sequences that minimize the number of substitutions required to explain the sequences observed at the tips of the tree. It uses information related to the grouping of taxa in the tree while amounts of evolution (i.e., the length of edges in the tree) are ignored. The accuracy

of the ancestral sequences estimated with the parsimony approach has been well studied from a theoretical perspective (Fischer and Thatte, 2009; Gascuel and Steel, 2010; Yang et al., 2011) and using simulations (Zhang and Nei, 1997).

Yang et al. (1995) and Koshi and Goldstein (1996) were the first to infer ancestral sequences using the maximum likelihood approach under a phylogenetic model. The inference relies here on a two-step approach. A phylogenetic model (i.e., a tree topology, with edge lengths and substitution model parameters) is first fitted to the data. Finding the nucleotide (or amino-acid or codon) character with the highest (marginal) posterior probability at any internal node in a fixed phylogeny is then relatively straightforward. Pupko et al. (2000) later described an elegant dynamic programming algorithm for the inference of the combination of character states at all internal nodes that maximizes the joint posterior probability.

These methods focus on selecting the best state in the alphabet of nucleotides, amino-acids or codons, i.e., the alphabet the data generating process relies upon. Although it is perfectly sensible from a statistical viewpoint, it does not always reflect potential uncertainty in the inference. Indeed, multiple characters can have high probabilities and selecting only the best one obliterates available information about the variability of ancestral sequence estimates (Blanchet et al., 2017). Ignoring uncertainty in the reconstruction of ancestral sequences is indeed a serious limitation of current estimation techniques that has been known for a long time (Cunningham et al., 1998). Previous attempts to deal with uncertainty in ancestral sequence inference consisted in generating a subset of ancestral sequences randomly sampled from their posterior probabilities (Gaucher et al., 2008) or considering all possible character states (or only a subset of sub-optimal residues) at positions deemed ambiguous (Thomson et al., 2005; McKeown et al., 2014). Eick et al. (2017) compared these two approaches on three families of protein domains and showed that sampling from the posterior distribution of residues produced non-functional proteins in some cases. In practice, the software FastML for estimating ancestral sequences using the maximum likelihood principle (Ashkenazy et al., 2012) can be used to select the k most probable ancestral sequences at each node, where k is fixed by the user.

In the present study, we introduce a new statistical criterion, the minimum posterior expected error (MPEE), and test another one, based on decision theory and the Brier score. Both reveal uncertainty in the inference without compromising on the intelligibility of ancestral character reconstruction. We also introduce and test several other criteria which, unlike MPEE and Brier-based criteria, require setting tuning parameters prior to the data analysis. We focus on the inference of past DNA and protein sequences using simulated data. Our results indicate that accommodating for ambiguity in ancestral sequences using MPEE or Brier amounts to a better use of the available data compared to the traditional approach.

2 Notations

The multiple sequence alignment is noted as $\mathbf{d}$. Its length, i.e., the number of columns in the alignment is l. Each sequence in $\mathbf{d}$ is a vector of characters, each character being taken within the alphabet $\mathcal{A}$ of nucleotides, amino-acids, codons or any other well-defined discrete states in finite number. Let n be the cardinality of $\mathcal{A}$. In what follows, $\mathbf{d}^{(s)}$ corresponds to the s-th column of $\mathbf{d}$. Let τ denote the unrooted topology of the phylogeny under scrutiny and $\mathbf{e}$ the vector of edge lengths on that tree. We assume that the tree is binary so that there are $u-2$ internal vertices, where u is the number of tips. Let $\mathbf{x}$ denote the vector of internal nodes and x is one of these nodes. $v_1(x)$, $v_2(x)$ and $v_3(x)$ are the three nodes directly connected to x, i.e., its three direct neighbors. $A_x^{(s)}$ is the random variable giving the ancestral character observed at node x and site s.

Inferring ancestral characters generally relies on evaluating the conditional probability $\Pr(A_x^{(s)} = \cdot|\tau, \mathbf{e}, \mathbf{d}^{(s)})$, which, for a particular character state y, is expressed using Bayes' theorem as follows:

$$\Pr(A_x^{(s)} = y|\tau, \mathbf{e}, \mathbf{d}^{(s)}) \propto \Pr(\mathbf{d}^{(s)}|\tau, \mathbf{e}, A_x^{(s)} = y) \Pr(A_x^{(s)} = y), \quad (1)$$

where $\Pr(A_x^{(s)} = y)$ is taken as the equilibrium frequency of state y since the substitution process at each site of the alignment is modeled as a homogeneous Markov chain running along the phylogeny. $\Pr(\mathbf{d}^{(s)}|\tau, \mathbf{e}, A_x^{(s)} = y)$ is the likelihood of the model given that state y is observed at node x. Assuming that node x has three neighbors, as is the case if the phylogeny is a fully resolved unrooted tree, $\Pr(\mathbf{d}^{(s)}|\tau, \mathbf{e}, A_x^{(s)} = y)$ is then obtained as follows:

$$\begin{aligned}\Pr(\mathbf{d}^{(s)}|\tau, \mathbf{e}, A_x^{(s)} = y) &= \prod_{i=1}^{3} \sum_{z \in \mathcal{A}} \Pr(\mathbf{d}_{v_i(x)}^{(s)}|\tau_{v_i(x)}, \mathbf{e}_{v_i(x)}, A_{v_i(x)}^{(s)} = z) \\ &\times \Pr(A_{v_i(x)}^{(s)} = z|A_x^{(s)} = y, e_{v_i(x)}), \end{aligned} \quad (2)$$

where $\mathbf{d}_{v_i(x)}^{(s)}$ corresponds to the part of $\mathbf{d}$ made of sequences found at the tips of the subtree rooted by $v_i(x)$ (hence $\mathbf{d}^{(s)}$ is the union of $\mathbf{d}_{v_1(x)}^{(s)}$, $\mathbf{d}_{v_2(x)}^{(s)}$ and $\mathbf{d}_{v_3(x)}^{(s)}$). $\tau_{v_i(x)}$ is the topology of this rooted subtree and $\mathbf{e}_{v_i(x)}$ are the lengths of its edges. $\Pr(A_{v_i(x)}^{(s)} = \cdot|A_x^{(s)} = \cdot, e_{v_i(x)})$ denotes the transition probability along the edge between x and $v_i(x)$ of length $e_{v_i(x)}$.

3 Inferring ancestral states

In the following two sections, we first introduce the standard criterion, i.e., the maximum a posteriori criterion, defined in the context of an unrooted or a rooted tree. The new, minimum posterior expected error and Brier-based criteria are presented next followed by other, less standard, approaches.

The maximum a posteriori (MAP) criterion

The most popular technique to infer ancestral states relies on the maximum a posteriori probability (MAP) criterion based on the marginal conditional probabilities defined above. The inferred character $\hat{y}$ is selected as follows:

$$\hat{y} = \underset{y}{\operatorname{argmax}}(p_y), \quad (3)$$

whereby $p_y \equiv \Pr(A_x^{(s)} = y|\tau, \mathbf{e}, \mathbf{d}^{(s)})$. This technique thus considers each node separately, selecting the most probable state at each of these nodes without any reference to the ancestral states inferred in other parts of the tree. As mentioned earlier, dynamic programming can also be used to infer the combination of ancestral states at all internal nodes that maximizes their joint posterior probability (Pupko et al., 2000). In practice however, many phylogenetic software rely on the marginal probabilities although PAML4 (Yang, 2007) and HyPhy (Pond and Muse, 2005) also return joint posterior estimates whenever the selected substitution model ignores the variation of substitution rates across sites, while FastML (Ashkenazy et al., 2012) implements a branch-and-bound algorithm to accommodate for this variability when building joint estimates (Pupko et al., 2002). Simulations suggest that the performance of the the marginal and joint approaches are virtually identical (Gascuel and Steel, 2014).

In Equation 3, the marginal posterior probability derives from the conditional likelihood as defined in Equation 2. Since the product in Equation 2 is over the three vertices directly connected to the internal node under scrutiny, the whole set of characters found at the tips of the

tree is involved in the calculation of this marginal probability. A distinct approach, which applies to the case where the tree is rooted, is to sum over the two vertices subtending the node under scrutiny (i.e., the two nodes "away from the root", taking the node of interest as reference). This solution amounts to considering that only the data "below" the node of interest convey information about the ancestral state. This reasoning is somehow at odds with the time-reversible markovian assumption about the substitution process as the position of the root is not identifiable under this class of models. More importantly, this approach unnecessarily ignores part of the data. It is nonetheless implemented in the software RAxML (Stamatakis, 2006). In the following, we will refer to this criterion as MAP_r.

The minimum posterior expected error (MPEE) criterion

The rationale behind MPEE rests on the definition of the loss function that the estimation of an ancestral character relies upon. The simplest way to define such function is as follows:

$$e_k(x,y) = \begin{cases} \alpha_k & \text{if } x \cap y \neq \varnothing,\ x \in \mathcal{A} \text{ and } y \in \mathcal{P}^+(\mathcal{A}) \\ \beta_k & \text{if } x \cap y = \varnothing,\ x \in \mathcal{A} \text{ and } y \in \mathcal{P}^+(\mathcal{A}), \end{cases} \tag{4}$$

where $e_k(x,y)$ is the error score when comparing x, the true character, and y the inferred one. We assume that $\beta_k \geq \alpha_k$ so that the cost of a mismatch between states x and y (i.e., $x \cap y = \varnothing$) is always greater or equal to that of a match (i.e., $x \cap y \neq \varnothing$). $\mathcal{P}^+(\mathcal{A})$ is the powerset of $\mathcal{A}$ minus the empty set. For instance, when considering binary character states, one may have $\mathcal{A} = \{\{0\},\{1\}\}$ and $\mathcal{P}^+(\mathcal{A}) = \{\{0\},\{1\},\{0,1\}\}$. We will refer to any element of $\mathcal{A}$ as a "character" while any element of $\mathcal{P}^+(\mathcal{A})$ will be referred to as a "character set". k is the level of ambiguity associated to y. It is equal to $|y|$, i.e., the number of elements in the character set y. For instance, when considering nucleotide sequences this time, if $y = \{A,C,G\}$, then $k = 3$ and the level of ambiguity is here equal to three.

For a given ambiguity level k, the posterior expected error associated to an inferred character set y is a weighted average of these errors. It is obtained as follows:

$$\mathrm{E}_k(y) = \sum_{x \in \mathcal{A}} p_x e_k(x,y), \tag{5}$$

i.e., for a given inferred character set $y \in \mathcal{P}^+(\mathcal{A})$, one computes the weighted average of the errors over all the potential true states $x \in \mathcal{A}$, where the weights are the posterior probabilities of x. Plugging the error scores (*sensu* Equation 4) into Equation 5, one obtains:

$$\begin{aligned} \mathrm{E}_k(y) &= \sum_{\substack{x \in \mathcal{A}, \\ x \cap y \neq \varnothing}} p_x \alpha_k + \sum_{\substack{x \in \mathcal{A}, \\ x \cap y = \varnothing}} p_x \beta_k & (6) \\ &= \alpha_k + (\beta_k - \alpha_k)(1 - \sum_{\substack{x \in \mathcal{A}, \\ x \cap y \neq \varnothing}} p_x), & (7) \end{aligned}$$

and the minimum posterior expected error for a given ambiguity level k, noted as $\mathcal{E}_k$, is thus:

$$\mathcal{E}_k = \alpha_k + (\beta_k - \alpha_k)\mathcal{P}_k, \tag{8}$$

where $\mathcal{P}_k = 1 - \sum_{i=1}^{k} p_{(i)}$ and $p_{(1)} \geq \ldots \geq p_{(n)}$ are the posterior probabilities of states in $\mathcal{A}$, ranked in decreasing order. Finally, the inferred ancestral character set, $\hat{y}$, is selected as follows: (1) find $\hat{k} = \mathrm{argmin}_k(\mathcal{E}_k)$, i.e., the optimal level of ambiguity, (2) given $\hat{k}$, return the subset of $\hat{k}$ non-ambiguous states with the $\hat{k}$ largest posterior probabilities. This subset, chosen in $\mathcal{P}^+(\mathcal{A})$, is the inferred character set $\hat{y}$. Hence, given values of α_k and β_k for $1 \leq k \leq n$, the time complexity for finding the ancestral state that minimizes the posterior expected error is $\mathcal{O}(n\log(n))$, i.e., the complexity involved in sorting the posterior probabilities, despite the alphabet considered (i.e., $\mathcal{P}^+(\mathcal{A})$) being of size $\mathcal{O}(2^n)$.

In order to define sensible ranges of values for α_k and β_k, we focus on the prior expected error:

$$\mathrm{E}_k^*(y) = \frac{1}{n} \sum_{x \in \mathcal{A}} e_k(x,y). \tag{9}$$

Our rationale here is that this prior error should be equal for every character set $y \in \mathcal{P}^+(\mathcal{A})$, i.e., a particular character set, no matter what its ambiguity level is, should not be more (or less) probable than any other character set before one actually observes and analyses the data available. Moreover, we choose to have $\alpha_1 = 0$ and $\beta_1 = 1$. This choice is not arbitrary. In fact, when the inferred state is chosen among $\mathcal{A}$ instead of $\mathcal{P}^+(\mathcal{A})$, setting $\alpha_1 = 0$ and $\beta_1 = 1$ leads to $\mathrm{E}_y = 1 - p_y$ and selecting the state y that minimizes the posterior expected error is therefore equivalent to applying the MAP criterion. Having $\alpha_1 = 0$ and $\beta_1 = 1$ makes the prior expected error equal to $(n-1)/n$ for all ambiguity levels. For a level of ambiguity k, the prior expected error is also defined as $(\frac{k}{n})\alpha_k + (\frac{n-k}{n})\beta_k$ by Equation 9, and the following equality thus holds:

$$k\alpha_k + (n-k)\beta_k = n - 1, \tag{10}$$

which defines a linear relationship between α_k and β_k. For any ambiguity level k, the minimum value that α_k can take is 0 as $\alpha_1 = 0$ and $\alpha_k \geq \alpha_1$ for all $k > 1$ (i.e., the loss when inferring y such that $y \cap x \neq \varnothing$ is greater or equal to that obtained with y', where $y' \cap x \neq \varnothing$, if $|y'| \leq |y|$). Its maximum is $\frac{k-1}{k}$, as derived from Equation 10 when setting $\beta_k = 1$ and $\beta_k \geq \alpha_k$. Having $\beta_k = 1$ as the minimum value that this parameter can take for any k is, here again, not arbitrary. Indeed, there is no good reason for the cost of an error when inferring an ambiguous character to be smaller than that of an error made when (wrongly) selecting a non-ambiguous character.

Replacing β_k in Equation 7 by the expression of that parameter as defined by Equation 10, the minimum posterior error for the ambiguity level k is then as given below:

$$\mathcal{E}_k = \begin{cases} \alpha_k \left(\frac{n-k-n\mathcal{P}_k}{n-k}\right) + \frac{(n-1)\mathcal{P}_k}{n-k} & \text{for } k = 1,\ldots,n-1 \\ 1 - \frac{1}{n} & \text{for } k = n. \end{cases} \tag{11}$$

The maximum value that $\mathcal{E}_k$ can take for any $k = 1,\ldots,n-1$ is obtained when $\alpha_k = (k-1)/k$ and $\mathcal{P}_k$ is maximum, i.e., $\mathcal{P}_k = 1 - \frac{k}{n}$, in which case $\mathcal{E}_k = 1 - \frac{1}{n}$. Therefore, $\mathcal{E}_k \leq \mathcal{E}_n$ and the fully ambiguous character set $\mathcal{A}$ is never strictly optimal (it is only optimal in the trivial case where $p_x = \frac{1}{n}$ for all x, in which case $\mathcal{E}_k = \mathcal{E}_n$ and all character sets are also optimal).

Because there is no obvious reason to set α_k to a particular value (apart for the cases where $k = 1$ and $k = n$), a natural approach is to integrate over all possible values for these parameters. Ideally one would thus consider all combinations of values $\alpha_1 \leq \alpha_2 \leq \ldots \leq \alpha_{n-1}$, evaluate $\mathrm{argmin}_{\{1,\ldots,n-1\}}(\mathcal{E}_1,\ldots,\mathcal{E}_{n-1})$ for these combinations and, together with $p_{(1)},\ldots,p_{(n)}$, identify the corresponding best character sets (one character set per combination). The character set inferred in the end would then be the one with the highest frequency among all combinations. Because we cannot accommodate for an infinite number of combinations in practice, we considered a finite grid of values for $\alpha_1 \leq \alpha_2 \leq \ldots \leq \alpha_{n-1}$. Furthermore, the procedure described above for identifying the best scoring character set would involve $\mathcal{O}(M^{n-1})$ operations, where M is the number of values of α_k considered (for any k) on the grid. In order to reduce this complexity, we considered a one-dimensional grid instead whereby each point on that grid defines values

of $\alpha_1, \ldots, \alpha_{n-1}$. The value of α_k on the $(i+1)$-th point on that grid, noted as $\alpha_k^{(i)}$, is calculated as follows:

$$\alpha_k^{(i)} = \left(\frac{k-1}{k}\right)\left(\frac{i}{M}\right), \quad i = 0, \ldots, M.$$

We then evaluate $k^{(i)} := \text{argmin}_{\{1,\ldots,n-1\}}(\mathcal{E}_1^{(i)}, \ldots, \mathcal{E}_{n-1}^{(i)})$, i.e., the ambiguity level that minimizes the posterior expected error at the $(i+1)$-th point on our grid of α_k values. Given $k^{(i)}$ and $p_{(1)}, \ldots, p_{(n)}$, we then identify $y^{(i)}$, the best scoring character set for each value of i. The inferred character set is finally obtained by selecting the element with the highest multiplicity in the multiset $\{y^{(i)}, i = 0, \ldots, M\}$. The computational complexity of this procedure is $\mathcal{O}(Mn)$, thereby making it considerably faster than the "full grid" approach described before.

Brier-based score

The MPEE criterion is not the only one that can be used to achieve the same goal of inferring ambiguous or non-ambiguous ancestral states. Very recently, Ishikawa and colleagues (2018) described an ancestral sequence reconstruction method based on the Brier scoring rule. They proposed to compare the posterior probabilities of non-ambiguous states to a null (or expected) uniform distribution on $1, 2, \ldots, n$ states. A squared Euclidean distance characterizes the difference between the observed and each of the n expected distributions. The inferred state, which can be ambiguous or not, is the one that minimizes these distances.

More specifically, for each k from 1 to n, the Brier-based (Brier-b) criterion is evaluated as follows:

$$\mathcal{B}_k = \sum_{i=1}^{k}\left(p_{(i)} - \frac{1}{k}\right)^2 + \sum_{i=k+1}^{n} p_{(i)}^2. \tag{12}$$

As with the MPEE criterion, one then finds $\hat{k} = \text{argmin}_k(\mathcal{B}_k)$, i.e., the optimal level of ambiguity, and the inferred ancestral state is that corresponding to the subset of $\hat{k}$ non-ambiguous states with the $\hat{k}$ largest posterior probabilities. The time complexity for finding the ancestral state that minimizes Brier-b criterion is $\mathcal{O}(n^2)$, making its calculation very quick in practice and applicable to both nucleotide and amino-acid character states. The corresponding method is called MPPA (Marginal Posterior Probability Approximation) in Ishikawa et al. (2018).

The score as defined above is inspired by the Brier scoring scheme which was originally designed to measure the accuracy of probabilistic predictions. In its original formulation, this score aims at evaluating how close a probabilistic prediction is from the actual outcome. It is obtained as follows:

$$\text{Brier} = \sum_{i=1}^{n}(p_i - o_i)^2, \tag{13}$$

whereby p_i is the probability that outcome i occurs as given by the prediction and $o_i = 1$ if event i actually occurred and $o_i = 0$ otherwise. This metric has sound statistical properties. Indeed, it is a proper scoring scheme (Murphy and Epstein, 1967): the probabilistic prediction distribution that minimizes the expected Brier score is the posterior distribution of character states under the model that generated the data. Following Ishikawa et al. (2018), we used the Brier score as defined in Equation 13 to compare the performance of the various ancestral inference criteria considered in this study.

Other criteria

The last two criteria presented above (MPEE and Brier-b) do not rely on any tuning parameter for inferring ancestral states, ambiguous or not. There are other criteria that achieve the same goal but involve tuning parameter(s) that are generally set in an arbitrary manner before starting the data analysis. We list below some of these criteria which performance were evaluated in the present study:

- Threshold criterion (Thresh): a threshold for the posterior probability of any non-ambiguous character set is defined *a priori*. Any character with a corresponding posterior probability greater or equal to that threshold is considered as valid, i.e., it belongs to the set of possible ancestral states. In our study, these thresholds were set to 1/4 and 1/20 when inferring ancestral nucleotides and amino-acids respectively.
- The cumulative probability criterion (CumProb): the (non-ambiguous) character states are first ranked in decreasing order of their posterior probabilities. The list obtained is then traversed from top to bottom. Character states keep on being added to the set of possible ancestral states as long as the corresponding cumulative probability does not exceed a certain threshold fixed *a priori*. In the present study, the threshold was fixed to 0.9.
- The differential criterion (Diff): a ranked list of non-ambiguous character states identical to that used by the CumProb criterion is first built. Characters keep on being added to the set of possible ancestral states as long as the difference between two successive posterior probabilities is less than a given threshold. In our study, the thresholds were set to 1/4 and 1/20 when inferring ancestral nucleotides and amino-acids respectively.

4 Simulations

The performance of the criteria introduced above was assessed using multiple simulated data sets. Each such data set consists in a phylogenetic tree along which sequences are evolved according to a standard Markov model of evolution. The specifics of these simulations are given below.

Simulating phylogenies

Random trees were generated with the R package *TreeSim* (Stadler, 2010). *TreeSim* creates random trees under a constant rate birth-death process. The function *sim.bd.taxa.age* was used to generate trees with a fixed number of taxa and fixed time since the most recent common ancestor of the sampled species. The number of taxa in the generated tree was set to 50. The tree height parameter (parameter '`age`' in `sim.bd.taxa.age`), H, was randomly drawn for each tree in $U[0.1; 1]$. 50 trees were generated in total. The birth rate was fixed to 0.1 while the extinction rate was fixed at 0.5. The trees hence generated are ultrametric, i.e., clock-like. Edge lengths are interpreted here as amounts of codon substitutions per (codon) site with an average length equal to 0.08.

Simulating sequences

We used the software `INDELible` (Fletcher and Yang, 2009) to simulate codon sequences along the trees. `INDELible` evolves sequences under a probabilistic model of molecular evolution that accommodates for point substitutions. It also accounts for insertion and deletion events although we did not consider this option in the present study. We used a codon model (the M0 model) that assumes that every codon and every lineage in the phylogeny evolves under the same substitution process whereby the dN/dS and transition/transversion rate ratios are fixed to 0.5 and 2.5 respectively, as suggested in the example file provided with `INDELible`. Codon frequencies at equilibrium were all equal. Sequences of length 300 codons were generated.

Computer programs and parameter settings

For each simulated data set, we inferred ancestral sequences using the GTR nucleotide substitution model (Tavaré, 1986), thereby using a model distinct from the one that generated the sequences. The parameters of the GTR model were estimated using maximum likelihood under the "true", i.e., the simulated phylogeny. A discrete gamma distribution was used here to accommodate for the variability of substitution rates across single sites in the alignment, thereby taking into account (although imperfectly) the rate heterogeneity due to the structure of the genetic code.

We also reconstructed ancestral states from the protein sequences resulting from the simulated codon sequences translated into amino-acids using the standard genetic code. We (wrongly) assumed that amino-acid sequences evolved under the LG model of substitution model (Le and Gascuel, 2008). For both nucleotide and amino-acid ancestral reconstruction, sequences were inferred along the true tree topology so that it is straightforward to match ancestral sequences to the corresponding estimated ones.

PhyML (Guindon et al., 2010) and RAxML (Stamatakis, 2006) were used to reconstruct ancestral sequences under the GTR and LG models. While the tree topologies used for the estimation were set to the true ones and fixed throughout the analysis, edge lengths and substitution model parameters (for the GTR model) were optimized with each software prior to the ancestral sequence reconstruction.

5 Results

Figure 1 presents the performance of the six criteria for inferring ancestral states considered in this study. Although sequences were simulated under a stochastic process describing changes between codons, a model of substitutions between nucleotides was assumed for the ancestral reconstruction (see Simulations section), thereby ignoring the non-independence between individual columns in each alignment. For each nucleotide site in the alignment and each internal node in the phylogeny, an ancestral state was inferred using one of the six criteria and compared to the "true" (i.e., simulated) nucleotide. Only cases where the maximum posterior probability of any nucleotide were smaller than 0.95 were considered here in order to focus on examples where the inference is not obvious.

For each pair of barplots and each criterion, corresponding to a given level of ambiguity in the inferred states, the ratio between the number of incorrectly reconstructed states (the true state is not in the set of inferred states, barplots on the right of each pair) and the number of correct inferences (the true state is in the set of inferred states, barplots on the left) gives an indication about the accuracy of the criterion. This ratio is also given for each criterion as a percentage in the column separating each pair of barplots (see "% err"). Also, comparing the plots obtained for the different ambiguity levels provides information about the precision of the different approaches: precise criteria will mostly infer non-ambiguous states ("# states: 1" in the figure) while imprecise ones will return ambiguous characters ("# states: 2, 3 and 4"). The "% tot." figure given below each error percentage corresponds to the percentage of inferred states in the corresponding ambiguity level. For instance, 100% of the states inferred using MAP are non-ambiguous while only 68% of those inferred with the Thresh criterion are non-ambiguous.

The percentage of errors obtained with the various criteria when considering only non-ambiguous inferred ancestral nucleotides varies from 24% for MAP to 8% for CumProb. The other criteria (Thresh, Diff, Brier-b and MPEE) all behave roughly the same with 14 to 18% of errors when considering non-ambiguous inferred states and less than 5% for ambiguous states. All criteria vary however in terms of the precision of the inference. Apart from MAP which, by construction, always returns non-ambiguous

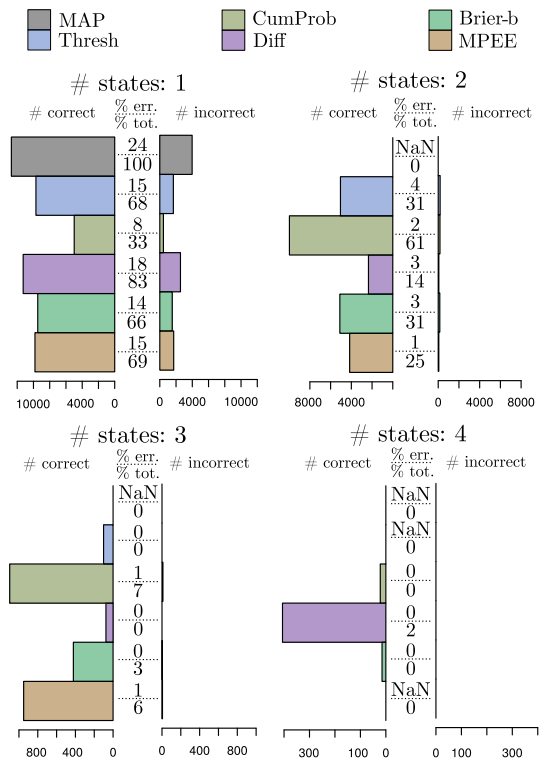

Fig. 1. Performance of several ancestral state reconstruction criteria applied to nucleotide sequences. The four plots correspond to the four ambiguity levels in the reconstructed nucleotide states, i.e., from one (the inferred ancestral state is not ambiguous) to four (the state is fully ambiguous). The left-hand (resp. right-hand) side of each plot gives the number of correctly (resp. incorrectly) reconstructed states across all internal nodes, all sites in each alignment and all alignments (although only cases where the highest posterior probability is smaller than 0.95 were examined). The column separating each pair of barplots for a given ambiguity level gives the percentage of errors (top) and the percentage of inferred states in the corresponding ambiguity level (bottom), separated by a dashed line, for each criterion. "NaN" stands for "Not a Number", resulting from divisions by zero. See text.

ancestral states, CumProb suffers from a lack of precision compared to other criteria (only 33% of inferences made with this criterion are non-ambiguous). The Diff criterion behaves well in terms of precision (83% of non-ambiguous), along with accuracy (18% of errors), this last figure being slightly greater than that of Brier-b and MPEE. Thresh, Brier-b and MPEE behave similarly in terms of precision and accuracy although the last two appear to be slightly more precise.

The results obtained for the reconstruction of ancestral amino-acid sequences are largely similar to that derived with nucleotides (Figure 2). Yet, it is worth noting that inferences performed with the Thresh criterion are seldom non-ambiguous (6% of non-ambiguous), which contrasts with the results obtained with nucleotide sequences (68% of non-ambiguous). Diff is, here again, not behaving as well as Brier-b or MPEE or any other criterion except MAP in terms of accuracy. Also, CumProb lacks precision with only 27% of all inferences being non-ambiguous. It is also interesting to note that Brier-b and MPEE perform very similarly, with a slight advantage for MPEE in terms of the percentage of errors when inferring ambiguous ancestral amino-acids (7 and 5% of errors with Brier-b for doubly- and triply-ambiguous inferences against 4 and 2% for MPEE). Brier-b score also appears to be slightly less precise compared to MPEE (39 *vs.* 31% of ambiguous characters inferred on amino-acid data with Brier-b *vs.* MPEE and 34 *vs.* 31% for nucleotide data).

We tested a range of values for the tuning parameters involved in the Diff, CumProb and Thresh criteria. With nucleotide data, setting the

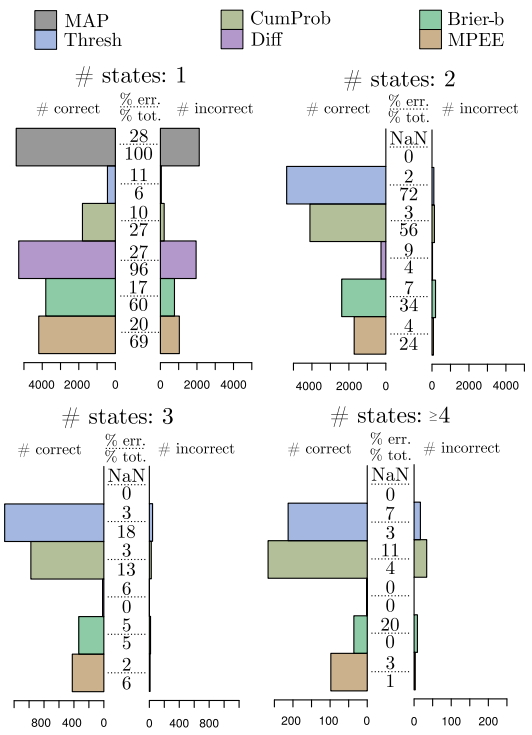

Fig. 2. Performance of several ancestral state reconstruction criteria applied to amino-acid sequences. Results obtained for ambiguity levels greater or equal to four (i.e., between four and twenty) were all grouped together and presented in the bottom-right corner. See caption of Figure 1 and text.

tuning parameters to more stringent values (0.05 for Diff, 0.5 for CumProb and 0.45 for Thresh), i.e., pushing these criteria towards selecting non-ambiguous states, forces the accuracy and the precision of these three approaches to become virtually identical to that of MAP (23%, 22% and 22% of errors for these three methods respectively, with 96%, 96% and 95% of non-ambiguous inferences). We also adjusted the tuning parameters of Diff, CumProb and Thresh so that the precision with these three criteria is close to that achieved by Brier-b and MPEE when considering non-ambiguous nucleotide states. The percentage of errors obtained with these criteria are then all very close to that obtained with Brier-b and MPEE (14% for Diff and CumProb, 15% for Thresh). This observation suggests that, considering non-ambiguous inferences only, it may not be possible to achieve a better accuracy than that obtained with Brier-b and MPEE for the precision obtained with these two techniques.

Finally, the percentage of errors obtained with MAP_r, i.e., the version of MAP that applies to rooted phylogenies and only considers characters below the internal node where the ancestral reconstruction is performed, are 46% and 49% with nucleotide and amino-acid data respectively. Compared to MAP (24% and 28% of errors), it is fairly obvious that ignoring part of the data is clearly detrimental and should be avoided.

In the present study, we used Ishikawa et al. (2018) Brier-based score as a tool to infer ancestral characters. Yet, the Brier score was originally proposed as a metric to measure the accuracy of probabilistic predictions. Following Ishikawa et al. (2018), we use this metric here to compare the performance of the various criteria considered in our study (see Equation 13). Results presented in Table 1 show that, for nucleotide sequences, the best methods are MPEE, Brier-b and Thresh, with scores fairly close to best (i.e., minimum) achievable score. The performance obtained with Diff and CumProb are slightly inferior while that of MAP is worse than Diff and CumProb. MAP_r is lagging far behind, with performance relatively close

Criterion	Nucleotides	Amino-acids
MPEE	0.37	0.45
Brier-b	0.37	0.44
Diff	0.40	0.54
CumProb	0.41	0.47
Thresh	0.37	0.54
MAP	0.48	0.56
MAP_r	0.93	0.97
Posterior	0.33	0.39
Random	1.50	1.90

Table 1. **Average Brier scores of the various inference criteria.** Averages were computed over the simulated data sets for which the highest posterior probability was smaller than 0.95. The second-to-last row gives the mean Brier score obtained when using the posterior probabilities of each nucleotide or amino-acid as predictive distribution, which is the minimum value the expected score can take (provided the posterior is evaluated under the model that generated the data). The last row gives the mean Brier score when predicted ancestral states are chosen uniformly at random, thereby defining an upper bound for the average score.

to that obtained with random predictions. Results obtained with amino-acid sequences are similar, although in this case, Thresh does not perform as well as what is observed with nucleotide data.

6 Discussion

This study focuses on new statistical criteria for the inference of ancestral nucleotide and amino-acid sequences. The main motivation behind our work was to improve the way uncertainty in the ancestral reconstruction is dealt with. In particular, in cases where multiple nucleotides or amino-acids have similar marginal posterior probabilities, selecting only one of them can be problematic. Doing so is prone to an increased amount of errors in the inference (see the accuracy of MAP in our simulations). Also, systematically selecting a single character state in the inference fails to convey important information about uncertainty in the decision process.

We test two classes of criteria here. One class has criteria that rely on tuning parameters that are set prior to the analysis (Thresh, CumProb and Diff). The others do not rely on any arbitrarily set parameter (MAP, Brier-b and MPEE). We assessed the performance of these techniques on sequences simulated under a codon model and analysed under nucleotide and amino-acid substitution models. Hence, in both cases, the models used for reconstructing ancestral states departed from the actual process that generated the sequences. While all criteria have strengths and weaknesses, depending on the type of data (nucleotides or amino-acids), the main goal of the analysis and the resources available (can one afford to consider multiple plausible ancestral states or only a single one will be taken into account?), Brier-b and MPEE behave well compared to other techniques. In fact, our results suggest that these two criteria are optimal in the sense that knowing the tuning parameter values for Thresh, CumProb or Diff that yield the same precision as that of Brier-b and MPEE, would not lead to higher accuracy. The MPEE criterion slightly outperforms Brier-b on our simulations, although the difference between the two in terms of both precision and accuracy is moderate.

The Brier-b and MPEE criteria for ancestral state reconstruction have a sound statistical basis. These approaches help unveil relevant information about the uncertainty in the inference in a concise format, which is relevant to biologists. Moreover, the computational overload associated to the application of these new criteria is negligible compared to the standard approach (i.e., MAP), thereby offering a practical alternative.

We thus recommend that these criteria are considered alongside the traditional approach for inferring the biochemical properties of ancient genetic sequences. In cases where reconstruction is ambiguous, as pointed out by Brier-b and/or MPEE, considering multiple alternative ancestral states could lead to a greater complexity in downstream analyses. Yet, this increase of complexity may also help unveil plausible properties of ancient molecules that would have been simply ignored otherwise.

7 Software availability

The MPEE and MAP criteria for inferring ancestral nucleotide and amino-acid states are implemented and documented in the PhyML software package (`https://github.com/stephaneguindon/phyml`). The scripts that were used to perform the simulations and analyze the results are available from `https://github.com/stephaneguindon/ancestral`.

Acknowledgments

We would like to thank Fabio Pardi for discussions along with Prof. Tal Pupko and two anonymous reviewers for their insightful suggestions that helped improve the original manuscript.

Funding

This research was supported by the Institut Français de Bioinformatique (RENABI-IFB, Investissements d'Avenir, ANR-11-INBS-0013) and the Agence Nationale pour la Recherche through the project GENOSPACE.

References

Ashkenazy, H., Penn, O., Doron-Faigenboim, A., Cohen, O., Cannarozzi, G., Zomer, O., and Pupko, T. (2012). FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Research*, 40(W1):W580–W584.

Blanchet, G., Alili, D., Protte, A., Upert, G., Gilles, N., Tepshi, L., Stura, E. A., Mourier, G., and Servent, D. (2017). Ancestral protein resurrection and engineering opportunities of the mamba aminergic toxins. *Scientific Reports*, 7(1):2701.

Bridgham, J. T., Carroll, S. M., and Thornton, J. W. (2006). Evolution of hormone-receptor complexity by molecular exploitation. *Science*, 312(5770):97–101.

Chang, B. S. and Donoghue, M. J. (2000). Recreating ancestral proteins. *Trends in Ecology & Evolution*, 15(3):109–114.

Cunningham, C. W., Omland, K. E., and Oakley, T. H. (1998). Reconstructing ancestral character states: a critical reappraisal. *Trends in Ecology & Evolution*, 13(9):361–366.

Eick, G. N., Bridgham, J. T., Anderson, D. P., Harms, M. J., and Thornton, J. W. (2017). Robustness of reconstructed ancestral protein functions to statistical uncertainty. *Molecular Biology and Evolution*, 34(2):247–261.

Fischer, M. and Thatte, B. D. (2009). Maximum parsimony on subsets of taxa. *Journal of Theoretical Biology*, 260(2):290–293.

Fitch, W. M. (1971). Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Zoology*, 20(4):406–416.

Fletcher, W. and Yang, Z. (2009). INDELible: a flexible simulator of biological sequence evolution. *Molecular Biology and Evolution*, 26(8):1879–1888.

Gascuel, O. and Steel, M. (2010). Inferring ancestral sequences in taxon-rich phylogenies. *Mathematical Biosciences*, 227(2):125–135.

Gascuel, O. and Steel, M. (2014). Predicting the ancestral character changes in a tree is typically easier than predicting the root state. *Systematic Biology*, 63(3):421–435.

Gaucher, E. A., Govindarajan, S., and Ganesh, O. K. (2008). Palaeotemperature trend for precambrian life inferred from resurrected proteins. *Nature*, 451(7179):704.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*, 59(3):307–321.

Ishikawa, S., Zhukova, A., Iwasaki, W., and Gascuel, O. (2018). A fast likelihood method to reconstruct and visualize ancestral scenarios. *bioRxiv*, page 379529.

Koshi, J. M. and Goldstein, R. A. (1996). Probabilistic reconstruction of ancestral protein sequences. *Journal of Molecular Evolution*, 42(2):313–320.

Le, S. Q. and Gascuel, O. (2008). An improved general amino acid replacement matrix. *Molecular Biology and Evolution*, 25(7):1307–1320.

McKeown, A. N., Bridgham, J. T., Anderson, D. W., Murphy, M. N., Ortlund, E. A., and Thornton, J. W. (2014). Evolution of DNA specificity in a transcription factor family produced a new gene regulatory module. *Cell*, 159(1):58–68.

Merkl, R. and Sterner, R. (2016). Ancestral protein reconstruction: techniques and applications. *Biological Chemistry*, 397(1):1–21.

Murphy, A. H. and Epstein, E. S. (1967). A note on probability forecasts and "hedging". *Journal of Applied Meteorology*, 6(6):1002–1004.

Pond, S. L. K. and Muse, S. V. (2005). HyPhy: hypothesis testing using phylogenies. In *Statistical Methods in Molecular Evolution*, pages 125–181. Springer.

Pupko, T., Pe'er, I., Hasegawa, M., Graur, D., and Friedman, N. (2002). A branch-and-bound algorithm for the inference of ancestral amino-acid sequences when the replacement rate varies among sites: Application to the evolution of five gene families. *Bioinformatics*, 18(8):1116–1123.

Pupko, T., Pe'er, I., Shamir, R., and Graur, D. (2000). A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Molecular Biology and Evolution*, 17(6):890–896.

Stadler, T. (2010). TreeSim in R-Simulating trees under the birth-death model. *R package*, 1.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, 17(2):57–86.

Thomson, J. M., Gaucher, E. A., Burgan, M. F., De Kee, D. W., Li, T., Aris, J. P., and Benner, S. A. (2005). Resurrecting ancestral alcohol dehydrogenases from yeast. *Nature Genetics*, 37(6):630.

Thornton, J. W. (2004). Resurrecting ancient genes: experimental analysis of extinct molecules. *Nature Reviews Genetics*, 5(5):366.

Yang, J., Li, J., Dong, L., and Grünewald, S. (2011). Analysis on the reconstruction accuracy of the fitch method for inferring ancestral states. *BMC Bioinformatics*, 12(1):18.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591.

Yang, Z., Kumar, S., and Nei, M. (1995). A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*, 141(4):1641–1650.

Zhang, J. and Nei, M. (1997). Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *Journal of Molecular Evolution*, 44(1):S139–S146.