Problems With Estimation of Ancestral Frequencies Under Stationary Models

EDWARD SUSKO1,2 AND ANDREW J. ROGER1,3

1Centre for Comparative Genomics and Evolutionary Bioinformatics Dalhousie University, Halifax, Nova Scotia; 2Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia; and 3Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia, Canada B3H 4H7

*Correspondence to be sent to: Edward Susko, Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, Canada B3H 3J5; E-mail: susko@mathstat.dal.ca.

Received 17 May 2012; reviews returned 14 August 2012; accepted 20 August 2012

Associate Editor: Mark Holder

Ancestral character state reconstruction is widely used in phylogenetics with widely varying goals. In many cases, reconstruction of character states is of interest at particular sites or subsets of sites. In other cases, the reconstructions are used to infer properties across a large number of sites such as frequencies of character states at an ancestral node. It is the latter problem that will be the focus here. As an illustration consider 2 studies, one by Gaucher et al. (2008) and the other by Di Giulio (2003), both of which were interested in the optimal growth temperature of the Last Universal Common Ancestor (LUCA). Gaucher et al. (2003) and Gaucher et al. (2008) reconstructed ancestral genes and synthesized these to determine optimal growth temperature. Di Giulio (2003) determined that a thermophily index, defined as a weighted average of amino acid frequencies, was consistent with a thermophilic or hyperthermophilic LUCA. The latter analysis is based on a demonstrated relationship between amino acid frequencies and optimal growth temperature which was further noted in Zeldovich et al. (2007). Thus, implicitly or explicitly, in the Gaucher et al. and Di Giulio studies, correct inference of thermophilic versus mesophilic depends heavily on the accuracy of the ancestral amino acid frequencies. The methods of Di Giulio (2003) use maximum a posteriori (MAP) estimates (Pagel 1999) coming out of a likelihood analysis and in Gaucher et al. (2008) they involved a weighting of characters by their posterior probabilities (WT).

Likelihood and Bayesian methods will be the focus of this article. However, maximum parsimony, because of its longer history of use and computational simplicity, has often been used for ancestral frequency reconstruction and has been used in a number of analyses that purport to establish substitution trends. For instance, for the pseudogenes considered in Gojobori et al. (1982), transitions were found to usually be more frequent than transversions, an exception being that the \( A \rightarrow G \) and \( T \rightarrow C \) substitutions were more frequent than their reciprocal substitutions. Such observations may be at least partially explained by the simulation results of Collins et al. (1994) and probabilistic analysis of Eyre-Walker (1998) which showed that parsimony reconstructions are more likely to produce ancestors where the most prevalent state among extant taxa occurs more frequently than it did among those extant taxa. More recently, Schmitz et al. (2000) used parsimony to conclude that nucleotide frequencies in primate mtDNA had changed over time. Because the more frequent nucleotides were the ones that were inferred to be more frequent in ancestors, the results are likely partly a consequence of biases of parsimony: Because parsimony is known to be statistically inconsistent for tree estimation, in contrast to likelihood and Bayesian methods, it may be expected that likelihood and Bayesian methods would not suffer from such difficulties unless models are misspecified. In commenting on Schmitz et al. (2000), Krishnan et al. (2004) established that MAP estimates showed a similar bias toward more prevalent extant taxa character states becoming even more frequent in ancestral sequences.

In ancestral reconstructions for which character state frequencies or frequencies of changes are of interest, it is implicit that the underlying model is nonstationary. Indeed, the reconstruction is often motivated by the possibility that ancestral frequencies will differ substantially from extant taxa frequencies. For example, the reconstructions of Gaucher et al. (2003), Di Giulio (2003), and Gaucher et al. (2008) were undertaken with the goal of determining whether the LUCA had proteins with amino acid frequencies, possibly different from any of the contemporary frequencies, that are consistent with those of a hyperthermophilic, thermophilic or mesophilic organism. Moreover, if indeed evolution is stationary throughout the tree, the frequencies of extant taxa should match those of ancestral taxa and thus provide appropriate estimates. Thus, the cases of
primary interest for ancestral frequency estimation are when evolutionary processes are nonstationary.

Although nonstationary generating models are of primary interest, the most commonly used software implementations for ancestral reconstruction (e.g., PAML, Yang 2007, 1997) assume stationary models. The likelihood methods that we consider and that are often used to obtain frequencies are MAP estimates (Pagel 1999) and the weighting of characters by their posterior probabilities (WT) (Boussau et al. 2008; Gaucher et al. 2008). Here, we critically assess the accuracy of such methods under non-stationary generating models. Because such ancestral reconstruction methods do take into account evolutionary divergence, it is not unreasonable to speculate that they will give better ancestral frequency reconstructions than simple use of extant sequence frequencies for nonstationary models. This turns out not to be the case. We focus on the estimation of ancestral frequencies by considering the limiting ancestral frequencies when methods are applied to data generated under simple nonstationary models. We show that all commonly used methods of ancestral reconstruction produce frequencies that do not converge upon the correct ancestral frequencies even with infinite sequence lengths. Weighting ancestral character states by their posterior probabilities produces estimates that are better than the frequencies of character states that had largest posterior probabilities. However, the resulting estimates are not much closer to ancestral frequencies than simply using extant sequence frequencies. The results apply for both Bayesian and maximum likelihood (ML) reconstruction methods when stationary Markov models are utilized. Because appropriate nonstationary models are capable of correctly estimating ancestral frequencies, it is important that nonstationary models be fit in cases where ancestral properties are of interest. Our analysis differs from the study of reconstruction methods in Williams et al. (2006) in that we find biases in frequency estimation even with simple nonstationary models that do not go away even with infinite data.

Although ancestral frequencies are often calculated under stationary models, there are some exceptions more recently as nonstationary models have been implemented. For instance, Boussau et al. (2008), the nonstationary model in nhPhyML (Boussau and Gouy 2006) was used for ML-based estimation and nhPhyloBayes (Blanquart and Lartillot 2008) was used in a Bayesian approach. An interesting hybrid method that we will consider here is that of Brooks et al. (2004) whereby ancestral frequencies were estimated through ML applied to a model that allowed these to be different from those of extant taxa. However, this was done after first fitting a homogeneous model to obtain edge lengths and other parameters.

Materials and Methods
To illustrate the properties of ancestral reconstruction, we considered the model of Galtier and Gouy (1998).
from the conditional distribution of character states for internal nodes given terminal node data. Such random generation gives rise to variation in results due to simulation: for the same data set, 2 different random generations will give rise to different ancestral frequencies. One way of adjusting for such simulation variation is to repeatedly generate sequences and average over generations. It turns out that the limiting frequencies, as the number of random generations gets large, can be calculated explicitly. Let \( x_i \) denote the terminal node data or site pattern for site \( s \). For instance, for the tree in Figure 1A, \( x_i = \text{AAC} \) gives a site pattern where the character states \( A \) are for terminal Nodes 1 and 2 but \( C \) is for Node 3. For an internal node of interest, let \( p(i|x_i; \xi) \) denote the probability that character state \( i \) is assigned to that node, given \( x_i \); here we explicitly indicate dependence on the parameters, \( \xi \), in the model which in the present setting include edge lengths and HKY model parameters. If random sequences are repeatedly generated to adjust for simulation variation, the long run frequency of state \( i \) (the frequency that the estimate converges upon as sequence length gets large) will converge upon \( p(i|x_i; \xi) \). Thus, averaging \( p(i|x_i; \xi) \) over sites gives the ancestral frequency estimate for the node.

The final set of methods we consider are described in Brooks et al. (2004) and will be referred to as quasi-ML (QML). Frequencies for the root node are of interest here. The method utilizes the EM algorithm (Dempster et al. 1977) to easily allow for different nonstationary model fitting and utilize the output (Dempster et al. 1977) to easily allow for different nonstationary model fitting and utilize the output of sophisticated stationary phylogenetic modeling software, all of the parameters other than the root frequencies are obtained from a stationary model fitting. With these parameters, the EM algorithm is applied to estimate the ancestral frequencies that maximize the likelihood.

The methods mentioned so far are all likelihood-based. Another major class of methods that are frequently used in practice are Bayesian methods (Huelsenbeck and Bollback 2001). Like the methods already discussed, Bayesian methods utilize the conditional probabilities of character states at internal nodes given terminal node data. However, rather than substituting estimates of edge lengths and other parameters into the expressions for these conditional probabilities, Bayesian methods treat parameters as random variables with prior distributions and calculate conditional probabilities averaging over parameter values. With this change in the manner by which conditional probabilities are calculated, MAP, JMAP, or WT ancestral frequencies can, in principle, be calculated as for likelihood methods. In practice, because of the difficulties of explicitly integrating over parameter values, Markov Chain Monte Carlo methods are used. What will be considered here, however, is the limiting ancestral frequencies of the different methods and it turns out that these are the same as for ML.

The Bayesian posterior for \( i \) given \( x_i \) can be expressed as

\[
p(i|x_i) = \frac{\int p(i|x_i; \xi) \exp[I(\xi)] \, d\xi}{\int \exp[I(\xi)] \, d\xi}.
\]

where \( I(\xi) \) is the log likelihood for the data and \( \pi(\xi) \) is the prior for \( \xi \). Separate Laplace approximations for the numerator and denominator in equation (1) can be obtained with suitable choices of \( I(\xi) \) and \( \pi(\xi) \) in Theorem 1 of Kass et al. (1995). For the numerator approximation, \( \beta(\xi) = -I(\xi)/n \), where \( n \) is the sequence length, and \( \beta(\xi) = \pi(i|x_i; \xi) \exp[I(\xi)] \, d\xi \). The denominator approximation is obtained by replacing \( \beta(\xi) \) with \( \beta(\xi) = \pi(\xi) \). Cancelling terms in the numerator and denominator and ignoring terms that are of higher order gives that as sequence length gets large, \( p(i|x_i) \approx p(i|x_i; \xi) \). Where \( \xi \) denotes the ML estimates of the parameters (see the Appendix). Thus, limiting Bayesian conditional probabilities of internal node data for Bayesian methods are the same as those for likelihood methods. It follows that the limiting ancestral frequencies coming from, for instance, WT ancestral character reconstructions will be the same regardless of whether Bayesian or ML methods are used. Thus, in what follows, we restrict attention to WT or MAP estimates from an ML analysis.

The limiting frequencies for the methods are obtained as a weighted average of the frequency contributions over all site patterns. The weights in averaging are the pattern probabilities for the generating model. Thus, for the weighting method (WT), the limiting frequency of character state \( i \)

\[
\sum_x p_x p(i|x; \xi).
\]

where \( p_x \) is the pattern probability for pattern \( x \) under the generating model and the sum is over all site patterns. For the methods that directly use character states like MAP, the limiting frequencies are of the form

\[
\sum_x p_x \beta_x(i).
\]

where \( \beta_x(i) = 1 \) or 0 according to whether the MAP ancestral character state for site pattern \( x \) is the state \( i \) or not.

All of the methods of reconstruction require estimates of the parameters for the model used in fitting. As is common practice, we assume that terminal node frequencies are used to estimate the stationary GC content. These will converge upon the terminal frequencies implied by the generating model and can be calculated from it. Assuming the other parameters, \( k \) and the edge lengths, \( t \) are estimated through ML, White (1982) established that under mild conditions, the estimates will converge upon the values that minimize the Kullback-Leibler divergence between them and the pattern probabilities under the generating model.
Specifically, in the present case, ML estimates of $\kappa$ and edge lengths $t$ will converge to the minimizer of

$$\sum_x p_x \log \left[ \frac{p_x}{p(x; t, \kappa)} \right].$$

(4)

The adjustment to the limiting probabilities for estimation is thus to use the parameters minimizing (4) in whatever ancestral reconstruction procedure is being considered. That is, MAP or WT estimates for a pattern are calculated with these parameters being substituted into the conditional probabilities of ancestral node character states.

The situation with QML is a little different. In this case, parameters of the stationary model are estimated first; these will converge upon the minimizers of equation (4). Then the EM algorithm is applied to estimate the model parameter ancestral GC content, $\omega$. It follows, from White (1982) that the QML estimate of ancestral GC content will converge upon the minimizer of

$$\sum_x p_x \log \left[ \frac{p_x}{p(x; t_{KL}, \omega_{KL}, \omega)} \right],$$

where $t_{KL}$ and $\omega_{KL}$ are the minimizers of equation (4).

RESULTS

We first consider results for the 3-taxa tree in Figure 1A when evolution is according to the nonhomogeneous model with $\omega = 0.2$, $\kappa = 2$, and various choices of $\theta$. Assuming the true parameters $(\omega, \kappa, \theta$, and $t$) are used for ancestral reconstruction, the long-run estimated ancestral GC content is plotted against $t$ in Figure 2. The ancestral reconstruction methods considered are MAP and WT. Also included are the long-run frequencies of GC content from the terminal node data (FREQ). Not indicated is QML which, in this case, with a single internal node and correct parameters gives limiting GC content equal to the ancestral GC content regardless of the values of $\theta$ and $t$.

The performance of WT and FREQ is qualitatively similar. Both give the correct ancestral GC content in the case $\theta = 0.2$, which corresponds to a stationary model. For other choices of $\theta$, using WT always gives limiting frequencies that are slightly closer to the correct ancestral frequencies than FREQ.

The MAP estimates consistently favor the character state that is most prevalent in the model. When the frequency of GC is less than AT $(\theta < 0.5)$, the long-run estimated ancestral GC is less than the correct ancestral value and when $\theta > 0.5$, the estimated ancestral GC is larger than the correct ancestral value. In fact, as $t$ gets arbitrarily large, the limiting MAP estimate of GC frequency converges to 1 if $\theta > 0.5$ and to 0 if $\theta < 0.5$. This is because for calculations with $t$ large, terminal node data are almost independent of ancestral data, and thus the probability of ancestral character state $i$ is approximately its stationary frequency $\pi_i$ for the model. Thus, for large enough $t$, the most likely character state will be the one with the largest $\pi_i$, regardless of what the site pattern was.

The limiting frequencies vary smoothly with $t$ for FREQ and WT but take jumps for MAP. This is because certain site patterns that for smaller $t$ almost always give the same ancestral reconstruction, suddenly give different ones. For instance, with $\theta = 0.4$ (Fig. 2B) when $t \leq 1.55$, ancestral state G (or C) is assigned whenever it appears 2 or more times at terminal nodes. The long-run frequencies are thus the sum of the pattern probabilities with either a G or a C appearing 2 or more times. This frequency varies continuously with $t$ since the pattern probabilities vary continuously. With $t = 1.56$, however, patterns like AGG no longer assign G to the ancestral node, giving rise to a sudden drop in the frequency with which a G or C is assigned.

Interestingly, for some parameter settings, and small to moderate edge lengths, MAP gives frequencies closer to the true frequencies than WT. However, unlike WT and FREQ, it fails to give correct frequencies when the model is correctly specified (Fig. 2A: the stationary model with $\theta = 0.2$) and is poor for larger edge lengths.

Adjusting for parameter estimation by substituting KL divergence minimizers, Figure 3 gives the results for the same settings as Figure 2. Although the WT frequencies are still closer to the generating frequencies than simply using terminal node frequencies, they are much more similar to terminal node frequencies. This makes intuitive sense, as the effect of parameter estimation is to select parameters that make the stationary model fit the nonstationary model as well as possible. In practice, frequencies are sometimes estimated through ML rather than the more usual choice of terminal node frequencies. We can adjust for the use of ML estimation of stationary frequencies in the calculation of limiting ancestral frequencies by adding the stationary frequencies to the parameters for KL divergence, rather than using the implied terminal node frequencies from the generating model. We did this and found that for the Figure 3 settings, the limiting estimated ancestral GC content was slightly closer to the actual ancestral GC content when ML estimates of stationary frequencies were used rather than terminal node frequencies. The differences between the 2, although small, was largest when time, $t$, was relatively small. A consequence is that the limiting WT and MAP frequencies using ML optimized frequencies differ only a little from those using terminal node frequencies.

Because the QML approach first estimates a stationary model and then estimates the ancestral frequencies, it no longer gives limiting frequencies that match the generating ancestral frequencies. However, a uniform, albeit, modest improvement in estimation is observed by comparison with WT across parameter settings.

Finally, we consider performance in a case where the ancestor is only indirectly related to extant taxa. In this case, since there are additional internal nodes, JTMAP can be included among the methods for ancestral frequency reconstruction. From Figure 4, we see that it consistently performs worse than MAP and that,
with the exception of short edges, neither MAP method performs well by comparison with the others.

The performance of WT and FREQ is almost identical, except for small edges. Although QML usually performs slightly better, surprisingly, QML performs worse than WT and FREQ when θ₀ is larger than θ₁ and t is large. The explanation may be that with large edge-lengths, WT and FREQ are unable to estimate the elevated stationary frequencies in the ancestral lineages, whereas QML estimates these without having enough signal to recognize these stationary frequencies were larger than the ancestral frequencies.

**DISCUSSION**

While ancestral frequency reconstruction has often been performed under homogeneous models, the fact that such reconstructions are desired in the first place suggests that a nonhomogeneous model is probably more appropriate; otherwise the frequencies of ancestral taxa can be expected to be similar to extant taxa. Our results indicate that, in such settings, MAP estimates of ancestral frequencies using ancestral reconstruction are likely to be very poor for moderate to large edge lengths.

Averages of conditional probabilities of character states consistently outperformed MAP estimates in the 6-taxon settings although they were sometimes not as good in the 3-taxon case when edge lengths were relatively short. Because WT and FREQ do not show as much of a tendency toward extreme and incorrect frequencies as MAP, they are desirable by comparison. In addition, in the case that the generating model is indeed stationary, weighted conditional probabilities converge upon the correct frequencies where MAP estimates do not. Yang (2006) Section 4.4.4 points out that one of the consequences is that MAP estimates will lead to an inference of composition changing over time when it is in fact stationary.

Use of the QML frequencies performed marginally better than average posterior probabilities except in the case of Figure 4C. Still, particularly for the average posterior probability, the gain over simply using extant taxa frequencies was not substantial. Although some methods did better than others at estimating frequencies, they usually did not converge upon the correct frequencies.

Figure 2. Long-run estimated ancestral GC content for the Galtier–Gouy model on the tree in Figure 1A, when the ancestral GC content, θ₀ = 0.2 and generating parameters are substituted as model parameters in estimation. Plotted are the limiting MAP estimates, extant taxa frequencies (FREQ) and weighted average posterior frequencies (WT). A) corresponds to the case that the generating model is stationary.
Williams et al. (2006) investigated ancestral reconstruction performance in the case of a complex nonstationary model that incorporated tertiary structure. Our results here are broadly congruent with those of Williams et al. (2006) and indicate that difficulties in ancestral frequency estimation persist even with infinite amounts of data and simpler models. (Note that what were referred to as ML and Bayesian estimates there are referred to as MAP and weighted (WT) here to distinguish them from fully Bayesian methods.) Similarly to Williams et al. (2006), we found that weighted estimates performed better than MAP, although usually only for longer edge lengths. Again, however, the gains were small in comparison with simply using extant taxa frequencies. Krishnan et al. (2004) compare MAP, WT, and parsimony reconstruction of ancestral frequencies. Comparisons were for real data (primate mitochondrial cytochrome b and cytochrome oxidase subunit I genes) and simulated data with generating parameters coming from the real data analysis. In real data analysis, MAP and parsimony frequency reconstructions were less similar to extant frequencies than WT reconstructions. Our results, which restrict attention to MAP and WT comparisons, are in general agreement with this observation. In simulations from a stationary homogeneous model, Krishnan et al. (2004) find that the biases in reconstruction tend to be larger for MAP, particularly for high-rate sites. Small sequence-length biases are not considered here. What we find, in addition, is that the biases of MAP do not go away even as sequence length gets arbitrarily large, whereas WT converges upon the correct frequencies for data generated under stationary models. This was illustrated in Figures 2A and 3A but holds more generally.

For nonstationary fitted models, as a consequence of the statistical consistency properties of ML estimation under a correct model (Wald 1949), the ML estimate of ancestral frequencies, which would be implied parameters in the model, will converge upon the correct frequencies. Alternatively, ancestral frequencies might be a byproduct of the complete ancestral reconstructions of WT or MAP. For WT, because the parameters of the model converge upon the generating parameters, the limiting frequencies are given by equation (3) which, because the $p_x$ in the numerator cancels with...
FIGURE 4. Long-run estimated ancestral GC content for the Galtier–Gouy model on the tree in Figure 1B, when the ancestral GC content, $\omega = 0.5$ and parameters are estimated through ML estimation. Plotted are the extant taxa frequencies (FREQ), weighted average posterior frequencies (WT), the QMLs and the limiting MAP estimates, both with joint maximization over all nodes (JTMAP) and maximization only at an ancestral node (MAP).

In conclusion, convergence upon correct ancestral frequencies in cases where ancestors and descendant compositions are different is likely only possible through implementation of nonstationary models, where ancestral compositions are estimated as parameters of the fitted model. For longer edge lengths, simple methods, like averaging the composition of MAP and JTMAP ancestral sequences obtained using stationary models, can seriously mislead researchers by providing values that appear to be different from extant frequencies but that are in reality drastic over- or underestimates of the true ancestral frequencies. Although a number of other methods like WT and QML have been implemented, their estimates hardly differ from the overall frequencies of extant taxa. By comparison, nonstationary models will converge upon the correct ancestral frequencies. It should be noted, however, that for long edges they too will have difficulties due to the larger variances in estimation at these edge lengths. For instance, considering the settings of Figure 2, with 1000 sites, we calculated the standard deviations of the ML
estimates of ancestral GC content under a nonstationary model to be at most 0.02 for edge lengths less than 0.5, at most 0.044 for an edge length of 1 but more than 0.19 for an edge length of 2. For such edge lengths, concatenated gene sets would clearly be needed to reduce variance. Nevertheless, methods based on stationary models would similarly have substantial variance at large edge lengths in addition to being biased toward extant taxa frequencies. Thus, the main implication of the inability of ancestral reconstruction methods to consistently estimate ancestral frequencies is that implementation of appropriate nonstationary models is the best bet for obtaining reliable estimates.

**Funding**

This research was supported by Discovery grants awarded to E.S. and A.J.R. by the Natural Sciences and Engineering Research Council of Canada. A.J.R. also acknowledges support from the Canada Research Chairs Program.

**Appendix**

**Laplace Approximation of Posterior Probabilities**

The Laplace approximation to \( p(\xi|x) \) in equation (1) obtained from Theorem 1 of Kass et al. (1990). Ignoring higher order terms, this theorem gives that, with some regularity conditions,

\[
\int b(\xi) \exp \left\{ -nb(\xi) \right\} dx = (2\pi)^{p/2} \left\{ \det \left( nh(\xi) \right) \right\}^{1/2} \exp \left\{ -nb(\hat{\xi}) \right\} p(\hat{\xi}) + O(n^{-1}) , \tag{A.1}
\]

where \( p \) is the dimension of the parameter vector \( \xi \). The numerator and denominator of equation (1) are of the same form as the left-hand side of equation (A.1) with \( h(\xi) = -c(\xi)/n \) in both numerator and denominator, but with \( h(\xi) = p(\xi|x_c) = p(\xi|\xi) \) for the numerator and \( h(\xi) = p(\xi|\xi) \) for the denominator. Therefore, in equation (A.1) cancel upon taking the ratio of equation (1), leaving only the ratio of the \( h(\xi) \) or \( p(\xi|x_c) = p(\xi|\xi) \).

There are conditions required for valid Laplace approximation. Because \( \xi \) is random, one wants these to hold with probability 1 as sequence length gets large. Simplified conditions for this are considered in Theorems 7 and 8 of Kass et al. (1990) and, for most standard phylogenetic models of sequence evolution, hold with the important additional condition that, in the generating model, edge lengths and frequencies are positive. Priors should also not be too informative. Sufficient conditions are that they be positive everywhere and remain bounded as well as having derivatives of second order; such conditions are satisfied by most priors used in practice.

REFERENCES


