Evaluating the robustness of phylogenetic methods to among-site variability in substitution processes

Mark T. Holder^{1,*}, Derrick J. Zwickl¹ and Christophe Dessimoz^{2,3}

¹*Department of Ecology and Evolutionary Biology, University of Kansas,
1200 Sunnyside Avenue, Lawrence, KS 66045, USA*

²*ETH Zurich, Institute of Computational Science, 8092 Zurich, Switzerland*

³*Swiss Institute of Bioinformatics, CMU Michel-Servet 1, 1211 Geneva, Switzerland*

Computer simulations provide a flexible method for assessing the power and robustness of phylogenetic inference methods. Unfortunately, simulated data are often obviously atypical of data encountered in studies of molecular evolution. Unrealistic simulations can lead to conclusions that are irrelevant to real-data analyses or can provide a biased view of which methods perform well. Here, we present a software tool designed to generate data under a complex codon model that allows each residue in the protein sequence to have a different set of equilibrium amino acid frequencies. The software can obtain maximum-likelihood estimates of the parameters of the Halpern and Bruno model from empirical data and a fixed tree; given an arbitrary tree and a fixed set of parameters, the software can then simulate artificial datasets. We present the results of a simulation experiment using randomly generated tree shapes and substitution parameters estimated from 1610 mammalian cytochrome *b* sequences. We tested tree inference at the amino acid, nucleotide and codon levels and under parsimony, maximum-likelihood, Bayesian and distance criteria (for a total of more than 650 analyses on each dataset). Based on these simulations, nucleotide-level analyses seem to be more accurate than amino acid and codon analyses. The performance of distance-based phylogenetic methods appears to be quite sensitive to the choice of model and the form of rate heterogeneity used. Further studies are needed to assess the generality of these conclusions. For example, fitting parameters of the Halpern Bruno model to sequences from other genes will reveal the extent to which our conclusions were influenced by the choice of cytochrome *b*. Incorporating codon bias and more sources heterogeneity into the simulator will be crucial to determining whether the current results are caused by a bias in the current simulation study in favour of nucleotide analyses.

Keywords: simulation; phylogenetic inference; codon model; mixture model; partitioned model; RY coding

1. INTRODUCTION

Many approaches to phylogenetic inference have been suggested (see Felsenstein 2003), so evaluating the relative merits of these methods has been an important aspect of phylogenetic theory for many years (cf. Kuhner & Felsenstein 1994; Huelsenbeck 1995; Hillis 1996). While the performance of methods on real data is the ideal way to compare methods, this approach is hampered by the fact that we rarely know the true tree. Evaluating the behaviour of methods on data generated by computer simulation provides an attractive alternative. However, with a few notable exceptions (e.g. Schoniger & von Haeseler 1995), the models of evolution assumed in the simulation studies are usually comparable in terms of complexity to the models assumed by inference methods.

A more stringent test of methods would generate data using models that are much more complex than those used for inference. Unfortunately, using complex models usually entails the specification of a large number of parameters. If reasonable values for these parameters are unavailable then the simulated data may be unrealistic, and we run the risk of conducting a challenging, but irrelevant experiment.

Here, we describe a data simulator designed to mimic the evolution of protein-coding DNA sequences, and present the results of a simulation study evaluating the accuracy of various phylogenetic reconstruction methods. The simulator implements the Halpern & Bruno (1998) model (HB model hereafter) of evolution. The model allows for site-specific amino acid frequencies; thus the chief challenge presented by the method is the substantial among-site variation in the pattern of evolution. As discussed by Ren *et al.* (2005), it is difficult or impossible to produce reasonable simulations of codon evolution from an amino acid or single-nucleotide simulator. Because the likelihoods calculated for DNA data cannot be meaningfully compared with that for the translated amino acid sequences, model choice that

* Author for correspondence (mtholder@ku.edu).

Electronic supplementary material is available at <http://dx.doi.org/10.1098/rstb.2008.0162> or via <http://journals.royalsociety.org>.

One contribution of 17 to a Discussion Meeting Issue 'Statistical and computational challenges in molecular phylogenetics and evolution'.

relies on likelihood ratios is not an option. A notable advantage of using a codon model of sequence evolution is that it is possible to compare the results of phylogenetic analysis conducted at the nucleotide, codon or amino acid level.

2. MATERIAL AND METHODS

Simulated data were generated under the HB model of evolution, which includes a separate set of equilibrium amino acid frequencies for each site in a protein-coding sequence. This model was formulated by considering a mutational process governing nucleotide substitutions. A set of site-specific, equilibrium amino acid frequencies affect the rate of non-synonymous codon changes. The model was originally described (Halpern & Bruno 1998) using the HKY model (Hasegawa *et al.* 1985) as the mutational model, but the extension to the general time-reversible (GTR) model (Lanave *et al.* 1984; Tavaré 1986) is trivial, and the simulations reported here use this GTR extension to the HB model.

The HB model is time reversible. The parameters of the model are the instantaneous rates of change, and transitions between codons that differ by more than one nucleotide are disallowed (this restriction is computationally convenient, but may be unrealistic; see Kosiol *et al.* 2007). For two codons, i and j , that differ by a single position, Halpern & Bruno (1998) defined, q_{ij} , the instantaneous rate of change from i to j as

$$q_{ij} = \frac{\phi_{n(j)} r_{n(i)n(j)} \ln \left(\frac{\pi_{a(j)} \phi_{n(i)}}{\pi_{a(i)} \phi_{n(j)}} \right)}{1 - \frac{\pi_{a(i)} \phi_{n(i)}}{\pi_{a(j)} \phi_{n(j)}}}, \quad (3.1)$$

where $n(i)$ denotes the nucleotide that codon i displays at the differing position; $a(i)$ indicates the amino acid encoded by codon i ; π represents the site-specific amino acid equilibrium frequency; ϕ is the base frequency parameter for the GTR process; and $r_{n(i)n(j)}$ is the symmetric factor of the GTR mutational process associated with a change in nucleotide from $n(i)$ to $n(j)$. By reversibility, $r_{n(i)n(j)} = r_{n(j)n(i)}$. If $\pi_{a(i)} \phi_{n(j)} = \pi_{a(j)} \phi_{n(i)}$ then q_{ij} is defined to be $\phi_{n(j)} r_{n(i)n(j)}$; thus the rate of synonymous changes is identical to the rate of these changes under a GTR model. The equilibrium frequency of a particular codon, ψ_i , in the model is simply

$$\psi_i = \pi_{a(i)} \frac{\phi_{n(i,1)} \phi_{n(i,2)} \phi_{n(i,3)}}{\sum_{j \in C_{a(i)}} (\phi_{n(j,1)} \phi_{n(j,2)} \phi_{n(j,3)})}, \quad (3.2)$$

where $C_{a(i)}$ is the set of codons that code for amino acid $a(i)$ and the notation $n(i,1)$, $n(i,2)$, $n(i,3)$ is used to indicate the nucleotides at the first, second and third base positions of codon i . As discussed by Halpern & Bruno (1998), the resulting model displays high rates of change between synonymous codons, with equilibrium amino acid frequencies determined by the set of π parameters inferred for each site.

(a) Parameter estimation

To use parameter values that are biologically plausible, the maximum-likelihood estimates (MLEs) of the parameters of the HB model were obtained from a dataset of 1610 cytochrome *b* sequences from mammals. Each of these sequences was unique (redundant sequences were culled) and was scored for at least 1000 sites. The cytochrome *b* gene is approximately 1140 nucleotides in most mammals. Alignment of the sequences for most of the length of the gene was non-problematic with most gaps being caused by incomplete sequences. Almost all of the length variation that

is not caused by incomplete sampling is the result of variation in the position of the stop codon. The end of the gene was cropped to avoid errors caused by misalignment in this region. The final alignment consisted of 1128 nucleotide positions (376 amino acids).

Source code for the program (written by M.T.H.) used for simulation and parameter inference is available from <http://people.ku.edu/~mtholder/bull>. For the cytochrome *b* alignment used here for parameter inference, there are 7152 free parameters in the model of substitution (7144 amino acid frequencies and 8 parameters of the GTR mutational process) and 3217 branch lengths to be estimated. Tree inference under the HB model was not attempted. Instead a tree was obtained from parsimony searches in which major recognized groupings among mammals were constrained to be monophyletic. A strategy of repeatedly performing constrained bootstrap searches and then constraining strongly supported groupings from these searches was employed to further constrain the tree in order to decrease search times. The final tree obtained is highly unlikely to be the most parsimonious tree given these constraints; however, previous workers (e.g. Yang *et al.* 1995; Sullivan *et al.* 2005) have noted that the MLEs of numerical parameters of the model of character evolution are often relatively insensitive to the choice of topology as long as the tree is a reasonably good one. Thus, the parameter estimates we obtained for the HB model may not be very different from the estimates that we would have obtained after conducting more rigorous tree searching.

Following tree estimation, MLEs of the branch lengths and parameters of the HKY + Γ_4 model were obtained from PAUP* v. 4.0.b08 (Swofford 2001). These parameters were used as initial values for the branch-length and mutational parameters of the HB model. The observed frequency of each amino acid at each site served as the basis for initial values for the amino acid equilibrium frequencies in the HB model. Maximization of the likelihood was performed by alternating between the rounds of substitution model optimization and branch-length optimization. This procedure was terminated when an entire round of optimization improved the likelihood by less than 0.05 log likelihood. This threshold was reached after 38 iterations. Branch lengths were optimized by recursively traversing the tree and applying the optimization procedure of Brent (1973) to each branch-length parameter. Parameters of the substitution model were optimized using the multidimensional parameter optimization scheme of Powell (1964) applied in batches with parameters for each amino acid position considered to be a group. The GTR mutation parameters were treated as a separate group of parameters for the Powell optimization. Amino acids that were not observed at a particular position in the data matrix frequently resulted in equilibrium-state frequencies of 0.0; this situation enabled the optimization of the likelihood calculation by recoding the character to an alphabet with fewer states. Despite the lack of observations of these states, non-zero parameter estimates for the associated amino acid frequencies were evaluated during parameter optimization, because it is possible that the MLE will contain a non-zero parameter estimate even for unobserved amino acids. This can occur because mutations affecting multiple sites are disallowed, and the codons of an unobserved amino acid can act as a 'mutational bridge' between sets of codons corresponding to amino acids that were observed. Each evaluation of non-zero frequencies required recoding the data to include states for the unobserved codons.

(b) Simulation of data

Because MLEs of the parameters of the HB model have been estimated only for the cytochrome *b* sequence dataset described above, a range of dataset sizes was created by simply concatenating multiple realizations of the cytochrome *b* sequence simulation. Sequence lengths of 1128 nucleotides and 11 280 nucleotides were simulated. These will be referred to as the $1\times$ and $10\times$ sequence lengths. Performance was evaluated on 100 simulated datasets. Trees of 50 leaves were generated in a multiple-step process. First, trees of 100 taxa were simulated using a Yule pure-birth model (Yule 1925). Next, random elimination of half of the leaves reduced the tree size to 50 taxa. Finally, deviations from ultrametricity were achieved by allowing the rate of evolution to change along the tree according to a constrained version of the relaxed clock model of Kishino *et al.* (2001). In this model, a rate of evolution is assigned to the root of the tree, and rates for descendant nodes are generated by drawing a random rate such that the difference in the logarithms of the rates at the beginning and end of a branch follows a normal distribution. The mean is chosen such that the expected rate is the same for the two nodes. The variance of the distribution is proportional to the branch length simulated in the Yule process. The model was constrained such that the minimum rate allowed for a node was one-tenth the maximum rate. Half of the model trees were simulated using high rates of sequence evolution to generate divergent sequences that approximate the tree lengths observed in this diverse set of mammalian cytochrome *b* sequences; these model trees are expected to be very difficult to reconstruct correctly. The mean tree length of these deep model trees was approximately 22 expected nucleotide changes per site. Pruning the 1610-taxa tree for the real cytochrome *b* sequences down to 50 sequences with representatives of the major groups results in a total tree length of approximately 25 expected changes per site. The other half of the model trees (referred to hereafter as the shallow trees) were generated using rates of sequence evolution that were lower by a factor of 10; the mean tree length of these trees was approximately 1.9 changes per site. See figures I and II in the electronic supplementary material for a depiction of four of the trees used. For each of these trees, five datasets of each of the two sequence lengths ($1\times$ and $10\times$) were simulated.

(c) Testing of inference methods

The focus of our study is to compare the performance of different models and criteria for selecting trees rather than evaluate the performance of heuristic searching methods. Thus, whenever feasible, we tested methods by starting tree searches from the tree used for the simulation; comparison with moderately extensive searches in which the true tree was not used as the starting tree shows that the results did not differ substantially (data not shown), indicating that the search space was relatively simple (but see the electronic supplementary material for discussion of the complexity of the space when very fast searches are used). Performance was evaluated by comparing returned trees to the model tree using the Robinson & Foulds (1981) (RF) distance between trees. During the simulation of each dataset, the sequence at the end of a branch was compared with the sequence at the ancestral node; if no changes were observed across a branch, then the branch was flagged for that replicate. For each simulated set of sequences, a tree that had all flagged branches collapsed was saved; returning this tree would represent detecting every internal branch that had any

phylogenetic signal during the simulation. Because synonymous changes would result in no signal for an amino acid-level analysis, an 'amino acid-collapsed' tree and a 'nucleotide-collapsed' tree were stored for each replicate. The performance of a method that always returned these trees is reported in the results to indicate the lowest RF that one could reasonably expect from an inference method.

(d) GARLI settings

ML searches were performed with the stochastic heuristic search software GARLI v. 0.96 Beta (Genetic Algorithm for Rapid Likelihood Inference; Zwickl 2006). Searches were started from the true tree used for simulation or from a tree generated by GARLI using an approximate ML stepwise-addition algorithm. All free model parameters were estimated for each search replicate. The automated stopping criterion was used, terminating each search after 10 000 generations without finding a significantly better scoring topology. The parameter controlling the precision of branch-length optimization was decreased five times over the course of each search. Other search settings were the defaults recommended in GARLI v. 0.96 Beta. For each dataset, model and starting point, two independent search replicates were performed, with the topology from the better scoring replicate saved.

GARLI searches were used to test the performance of ML inference on nucleotide and RY-coded (purines were recoded as R and pyrimidines as Y) data, amino acid sequences and codon-level analyses. For nucleotide and RY data, the GTR+I+ Γ_4 model consisting of the GTR substitution process (Lanave *et al.* 1984; Tavaré 1986) with a four-category discrete approximation to gamma-distributed rate heterogeneity (Yang 1994) and an inferred proportion of invariable sites was used. The empirical amino acid matrix of Whelan & Goldman (2001), with the equilibrium amino acid frequencies set to those observed in each alignment (+F), discrete gamma-distributed rate heterogeneity with four rate categories (+ Γ_4) and an inferred proportion of invariable sites (+I) was used for amino acid analyses. Codons were analysed using the Goldman & Yang (1994) model with a single non-synonymous/synonymous relative rate parameter as well as the discrete (M3) model of Yang *et al.* (2000), with three and four non-synonymous/synonymous relative rate categories. For all codon models, the equilibrium codon frequencies calculated by the F3 $\times$ 4 method (codon frequencies calculated as the product of the observed nucleotide frequencies at the three-codon positions in the alignment, normalized to account for stop codons).

(e) PHYLOBAYES

PHYLOBAYES v. 2.3 (Lartillot *et al.* 2004) was used to evaluate the CAT model for amino acid analysis (Lartillot & Philippe 2004) (referred to as 'aaBayes'). Runs were conducted until the maximum difference in split frequency between two runs was less than 0.1 and then terminated. Default values were used for all priors.

(f) MRBAYES settings

MRBAYES v. 3.2 (Ronquist & Huelsenbeck 2001) was used to test Bayesian inference under nucleotide models ('nucBayes' in the results). The GTR+I+ Γ_4 model was used with the following priors: branch lengths $\sim \text{Exp}(\text{mean}=0.1)$; proportion of invariant sites $\sim \text{Uniform}(0,1)$; base frequencies $\sim \text{Dirichlet}(1, 1, 1, 1)$; symmetric component of the instantaneous rate matrix $\sim \text{Dirichlet}(1, 1, 1, 1, 1, 1)$; and shape parameter for the gamma-distributed rates $\sim \text{Exp}(\text{mean}=1)$. At least three independent runs with four Metropolis-coupled

chains (Geyer 1991) were conducted, the sample frequencies were set to 500 and the automated stopping rule was used with a stop value of 0.05 (and 0.02 in the case of some runs that terminated quickly). The first half of the sampled trees were discarded as burn-in for the purpose of summarization. MRBAYES was used to conduct nucleotide analyses under the GTR+I+ Γ_4 model and a partitioned version of GTR+I+ Γ_4 in which a set of model parameters (including mean rate of evolution) were allowed for each of the codon positions. The settings used here correspond to a partitioned version of the priors and Markov chain Monte Carlo settings described above. A Dirichlet(1, 1, 1) distribution was used as the prior on the rate multiplier for each of the subsets in the partitioned analysis.

(g) Parsimony searches

PAUP* v. 4.0.b10 (Swofford 2001) was used for parsimony searches. Tree-bisection–reconnection searches were conducted starting from the true tree with a maximum of 10 equally parsimonious trees retained at any step. Performance was based on the first of the trees returned in those replicates that returned more than one tree.

(h) Distance-based searches

Distance-based searches under balanced minimum evolution (BME) (see Desper & Gascuel 2004), the Fitch & Margoliash (1967) form of least squares (LE) and minimum evolution (ME) were tested. FASTME v. 1.1 was used for searching under the balanced minimum-evolution criterion (see Desper & Gascuel 2004) using the ‘balanced NNI’ swapping and ordinary LE using both balanced NNI and OLS NNI swapping routines (Desper & Gascuel 2002). Other distance searches were conducted in PAUP*; in the case of minimum-evolution criteria, negative branch lengths were allowed but set to zero for tree-length calculation. Neighbour-joining (Saitou & Nei 1987) and BIONJ (Gascuel 1997) were tested. Searches starting from the true tree for both the minimum-evolution criterion (Rzhetsky & Nei 1992) and least-squares criterion were also studied. The weighting suggested by Fitch & Margoliash (1967) was applied during branch-length optimization. PAUP* was also used for distance corrections to create the input for FASTME (Desper & Gascuel 2002, see below).

The HB model does not employ gamma-distributed rate heterogeneity, so it is not possible to evaluate the performance of distance methods under the ‘true’ value of the shape parameter. Using the parameters fit to the cytochrome *b* data, only 90 out of the 1128 nucleotide positions (and 15 out of the 376 amino acid positions) were invariant, but many other sites experience very low rates of evolution. Previous workers (e.g. Takahashi & Nei 2000; Guindon & Gascuel 2002) have reported that correcting distances using a simpler model of evolution can lead to better performance relative to analyses that use distance corrections based on either the true model or the model preferred by model selection criteria such as the Akaike information criterion (AIC; Akaike 1973). To investigate these effects, we conducted inference under sweep over models of substitution (JC, K2P, HKY and GTR) with either invariant sites or gamma-distributed rates. Models using a proportion of invariant sites from 0.05 to 0.6 were studied (in steps of 0.05). Values of 0.1, 0.13, 0.15, 0.17, 0.2, 0.25, 0.3, 0.35, 0.4, 0.5, 0.6, 0.7, 0.9, 1.0, 1.2, 1.5 and 2.0 were used for the shape parameter of the gamma distribution. Corrected pairwise distances were calculated in PAUP* for these parameters using the JC (Jukes & Cantor 1969), K2P (Kimura 1980), HKY and GTR models. Distance corrections were also performed using values of the shape parameter

preferred by the GAME method of Guindon & Gascuel (2002) for the JC+ Γ , K2P+ Γ and TN93+ Γ (Tamura & Nei 1993) models calculated using both the FASTME and BIONJ options within GAME.

Other rate matrices and models which were examined include the F84 (Felsenstein 2004), WAG (Whelan & Goldman 2001), GCB (Gonnet *et al.* 1992) and CBRGDNA07 (an unpublished, empirical DNA rate matrix estimated from mammalian orthologues in OMA by Dessimoz *et al.* 2005). Pairwise distance estimation under F84 was done using dnadist from the PHYLIP package (Felsenstein 2004), while distances under WAG, GCB and CBRGDNA07 were estimated using an *ad hoc* script in the DARWIN environment (Gonnet *et al.* 2000). Weighted least-squares trees were inferred by the procedure ‘MinSquareTree’ from DARWIN (Gonnet *et al.* 2000). The weights are either the distance variance estimates (for distances from DARWIN) or the distance themselves (for distances from dnadist). A total of 656 combinations of distance-based analyses were employed on each dataset to verify that the relatively poor performance of distance methods was not simply the result of improperly choosing a distance.

3. RESULTS AND DISCUSSION

(a) Description of the simulation model

Parameter estimates from the mammalian cytochrome *b* revealed substantial base frequency bias (equilibrium frequencies for the GTR mutational process of 53.7, 26.5 and 3.9 per cent for A, C and G, respectively). Interactions with the amino acid frequencies in the model lead to very different base frequencies by codon position. This variation in base frequencies across codon position has been noted in cytochrome *b* sequences previously (see Voelker & Edwards (1998) for a discussion of the substantial variation in the parameters of nucleotide models across the different codon positions of cytochrome *b* in birds); in fact, the empirical base frequencies for the third base position in the mammalian cytochrome *b* dataset were 40.4, 34.9 and 3.2 per cent for A, C and G. As expected, the amino acid frequencies showed among-site heterogeneity with evidence of overfitting—each site in the protein resulted in parameter estimates in which at least one amino acid’s frequency was inferred to be zero, thereby prohibiting the occurrence of that amino acid in the simulations. We do not claim that the HB model or parameter estimates are perfect reflections of the evolutionary pressures on cytochrome *b*, so the overfitting may not be too troubling. Further studies on different genes, a wider range of taxa and explorations using pseudocounts for unobserved amino acids to avoid overfitting are warranted.

Across all sites in the inferred model, the mean number of codons with non-zero equilibrium frequencies was approximately 21, with a median number of allowed codons of 18. Twenty-two positions in the inferred model are constrained to have four or fewer codons (six positions were constrained to a twofold degenerate codon). The mean of the expected rate for first position sites is 2.7 times higher than the mean rate for second position sites. The mean rate of the third position sites is 9.6 times higher than the mean second position rate. All base positions exhibit some among-site rate heterogeneity (with less than 10% variation in rate between the fastest sites in each position). All of

Figure 1. Box plots of the RF distance from the true tree for different phylogenetic methods on the deep-1 $\times$ experiment. When an inference category included more than one tested method, we present the results for the variant (run settings) that yielded the lowest mean RF score for the combination of tree depth and sequence length shown in the figure. For each column, the box is circumscribed by quartiles, the median is a bold central line, whiskers denote the most extreme data (that are not outliers) and circles indicate outliers. Methods displayed are as follows: codon: ML searches in GARLI using the M3 or M3 $\times$ 4 codon model; aaColl: Model tree with all branches without amino acid changes collapsed; aaBayes: CAT mixture of Poisson models on amino acid data (PHYLOBAYES); aaML: WAG + I + Γ model (GARLI); nucColl: model tree with all branches without nucleotide changes collapsed; BME: FASTME searches (under the best-performing model for distance corrections); LS: searches under the least-squares criterion (Fitch–Margoliash weighting; under the best-performing model for distance corrections); ME: searches under the minimum-evolution criterion (under the best-performing model for distance corrections); Pars: parsimony searches (from the true tree) in PAUP*; nucML: ML searches in GARLI using the GTR + I + Γ model; nucBayes: Bayesian inference in MrBAYES using GTR + I + Γ . Unpartitioned analyses performed the best on the shallow tree and partitioned (by codon position) analyses performed the best on the deep tree simulations; RY: ML searches on RY-coded data in GARLI using the GTR + I + Γ model.

the third base position sites were variable and even the slowest third base position sites have a rate of evolution that is close to the mean rate taken across all positions (see figure V in the electronic supplementary material). At the first and second positions, some sites are fixed (40 first base position sites and 50 second base position sites). The relatively low amount of rate heterogeneity in the third base position of the HB model is one example of an unrealistic aspect of the model.

(b) Performance of phylogenetic inference methods

We surveyed a large number of methods because the primary goal of the study is to determine how robust different methods are. *A priori* we would expect some of the methods (e.g. the CAT model and the partitioned-by-codon-position analyses in MrBAYES) to be better suited for analysing the heterogeneous model. However, none of the inference methods examined here corresponds exactly to the true model; thus these expectations may not be met. The simulator allows us to examine the behaviour of inference procedures directly. Owing to the large number of analyses conducted (especially the large number of models used for distance corrections),

we will focus our discussion on the settings that yielded the best performance for each category of inference methods, where performance is judged by the RF distance from the returned tree to the true tree (see the electronic supplementary material for a full summary of the results). Figures 1–4 reveal that performance was quite variable. As expected, the error (as measured by the RF distance to the true tree) decreased when longer sequence lengths were examined. All methods did poorly at the task of reconstructing the trees from the deep tree simulations. This result is not too surprising, given the very long branches on the trees (with tree lengths of 15–22 expected changes per site for a 50 leaf tree), but the results do contrast with the more favourable results obtained by Hillis (1996) on simulations using simple simulation models on large trees. Interestingly, the RY-coding schemes improved performance of the GTR + I + Γ_4 model on the deep trees simulation with 10 $\times$ sequence length. This implies that at least some of the missed branches are the result of misinterpreting the phylogenetic signal rather than a lack of signal in these datasets.

In general, nucleotide-level analyses using ML or Bayesian approaches appear to outperform amino acid-level analyses, codon analyses and both distance and

Figure 2. Box plots of the RF distance from the true tree for different phylogenetic methods on the deep-10 $\times$ experiment. The other descriptions are the same as given in the legend of figure 1.

Figure 3. Box plots of the RF distance from the true tree for different phylogenetic methods on the shallow-1 $\times$ experiment. The other descriptions are the same as given in the legend of figure 1.

parsimony methods. However, the improvement of likelihood-based procedures over distance-based inference was slight in the case of the deep tree simulations. Although the simulation model itself is formulated at the codon level, it is perhaps not surprising that the codon-based analyses performed poorly. By only inferring dN/dS parameters that govern the overall rate of amino acid-changing substitutions, the GY94

model assumes that the relative rates of change between different types of codons are constant over the length of the sequence. This is in strong contrast with the HB model, which allows a strong site-specific preference for particular amino acids. It would be interesting to investigate the performance of methods, such as the PASSML model of Liò *et al.* (1998), which consider spatial structures within protein-coding sequences.

Figure 4. Box plots of the RF distance from the true tree for different phylogenetic methods on the shallow-10 $\times$ experiment. The other descriptions are the same as given in the legend of figure 1.

Longer sequence lengths were produced by concatenating several simulations of cytochrome *b*-like sequences. Given that the empirically derived rate matrices used for amino acid inference (such as the WAG matrix) have been derived from a large set of proteins, it is quite plausible that these methods would perform better in simulations that are capable of producing a more diverse set of test sequences. Thus, the current simulations may not be a fair assessment of the performance of amino acid-level analyses.

In the shallow tree simulations, Bayesian inference on nucleotide data performed slightly better than ML inference under the GTR+I+ Γ model. However, the comparison of the false-positive and false-negative rates of methods and the comparison of the MAP tree with the tree returned by a majority-rule consensus approach (see tables 1 and 2 in the electronic supplementary material) reveal that this slight performance advantage is largely the result of returning a tree with polytomies in majority-rule consensus trees from Bayesian analyses. While collapsing weakly supported branches is certainly a reasonable and well-justified practice (Berry & Gascuel 1996; Holder *et al.* 2008), analogous procedures for ML inference have not been investigated on these datasets (so this behaviour may be regarded as a bias in favour of Bayesian methods in this study). Use of the tree with the highest posterior probability results in higher mean RF distance to the true tree for the Bayesian analyses. Examining the bootstrap support for the nodes returned by ML methods would be an interesting future work with these simulations. Note that, at least in the case of longer sequences, the higher accuracy of amino acid inference using the CAT model compared with ML inference under the WAG+I+ Γ does not appear to be the result of

returning less resolved trees—the CAT model has a lower mean false-negative rate on the 10 $\times$ simulations.

Partitioned analyses on the nucleotide level (in MrBAYES) and the CAT-mixture model are designed to deal with the data that are generated by heterogeneous processes. The results here show some evidence that these techniques are effective. The CAT model outperformed the WAG+I+ Γ model on amino acid data; on the deep tree simulations, the partitioned GTR+I+ Γ inference in MrBAYES outperformed the unpartitioned GTR+I+ Γ (see the electronic supplementary material). Preliminary results from a small number of replicates (not shown) indicate that the GTR-CAT model for nucleotide analyses in PHYLOBAYES may be a very effective method for these data.

The choice of rate heterogeneity parameters for distance corrections had a larger effect than the model of evolution chosen (see the electronic supplementary material, figure III). Fortunately, automated methods (use of MLEs in the I+ Γ_4 model and the GAME approach for the + Γ_4 model) for choosing an appropriate level of rate heterogeneity did quite well on the shallow tree simulations. These methods exhibited performance close to the best accuracy encountered for any of the models found in the parameter sweep across alpha and the proportion of invariant sites (see table 1 in the electronic supplementary material). For the shallow tree simulations, using the + Γ model with the shape parameter determined by GAME was slightly better than using the I+ Γ model with the parameters determined by the MLEs from GARLI; however, the situation was reversed for the deep tree simulations (see table 2 in the electronic supplementary material). In agreement with Desper & Gascuel (2002, 2004), we found that

BME outperformed ME under most conditions (see figure IV in the electronic supplementary material).

The best performance for the minimum-evolution criterion on the 10× sequences lengths (see table 2 in the electronic supplementary material) for the deep tree simulations came from using the HKY model with an extremely high proportion of invariant sites ($I=0.6$). The RF to the true tree was very low for ME under this model (much lower on some datasets than any other method), but the variance in RF was very high. BIONJ (which otherwise performed quite well compared with searching under the minimum-evolution criterion) performed much worse on the same distance correction. We suspect that the results for ME under this model are an artefact of starting tree searching from the true tree, and that the best (non-artefactual) performance for ME on the 10×, deep tree simulations was actually from using JC+ Γ with $\alpha=0.1$; results from this distance correction were used to represent the performance of ME methods in figure 2. BIONJ performed slightly better than ME on this model.

4. CONCLUSIONS

The development of realistically complex data simulators would represent a significant advance in our ability to evaluate proposed methods of phylogenetic inference. We believe that the HB model used here has many of the features that such a simulator should display—substantial among-site heterogeneity and inclusion of the effects of both mutation and selection. However, the results presented here on the relative performance of different styles of phylogenetic analysis should be viewed with caution. There are a large number of caveats to bear in mind with the simulator and parameters presented here. Parameters have been estimated from one gene only. Furthermore, the reliability of cytochrome *b* sequences for deep phylogenetic problems is questionable (see Naylor & Brown (1998) for the performance of mitochondrial genes in reconstructing a commonly accepted tree). This implies that these simulations may be underestimating the strengths of some analyses (e.g. codon- and amino acid-level analyses) that may perform well on deep phylogenetic inference from a diverse set of protein-coding sequences.

The mutational process is assumed to be constant across all sites in the HB model. This assumption is not well justified and may result in a bias in favour of nucleotide-based models by underestimating the heterogeneity of the process of nucleotide evolution. The lack of codon bias may also represent a missing factor that could confound nucleotide-level inference. Thus, more realistic simulations might cause the performance of nucleotide-level inference to decrease, while sampling from a wider set of genes might improve the performance of a model such as the WAG model that was trained on a large set of gene sequences. For this reason, it is unclear whether the relatively poor performance of the WAG model will be a general result from complex simulation studies. The assumption of homogeneity of process across the tree is also a weakness of this approach, though it is not obvious

whether any of the inference methods tested were favoured by failing to simulate non-stationary evolution. While Martin & Palumbi (1993) found strong conservation of the kinds of amino acid substitutions in cytochrome *b* between sharks and mammals, estimating parameters of the HB model from multiple clades may provide a sense of the extent to which the amino acid frequencies do change.

The authors would like to thank David Hillis, Junhyong Kim, Tracy Heath, Ziheng Yang, Nick Goldman and two anonymous reviewers for their helpful discussions and comments. M.T.H. thanks NSF for financial support (grants EF 03-31495 to David Swofford and DEB-0732920 to M.T.H.).

REFERENCES

- Akaike, H. 1973 Information theory as an extension of the maximum likelihood principle. In *Proc. 2nd Int. Symp. on Information Theory* (eds B. N. Petrov & F. Csaki), pp. 267–281. Budapest, Hungary: Akademiai Kiado.
- Berry, V. & Gascuel, O. 1996 On the interpretation of bootstrap trees: appropriate threshold of clade selection and induced gain. *Mol. Biol. Evol.* **13**, 999–1011.
- Brent, R. P. 1973 *Algorithms for minimization without derivatives*. Englewood Cliffs, NJ: Prentice-Hall.
- Desper, R. & Gascuel, O. 2002 Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J. Comput. Biol.* **9**, 687–705. (doi:10.1089/106652702761034136)
- Desper, R. & Gascuel, O. 2004 Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. *Mol. Biol. Evol.* **21**, 587–598. (doi:10.1093/molbev/msh049)
- Dessimoz, C., Cannarozzi, G., Gil, M., Margadant, D., Roth, A., Schneider, A. & Gonnet, G. 2005 OMA, a comprehensive, automated project for the identification of orthologs from complete genome data: introduction and first achievements. In *Proc. RECOMB 2005 Workshop on Comparative Genomics* (eds A. McLysath & D. H. Huson). Lecture Notes in Bioinformatics, pp. 61–72. Berlin, Germany: Springer.
- Felsenstein, J. 2003 *Inferring phylogenies*. Sunderland, MA: Sinauer Associates.
- Felsenstein, J. 2004 *PHYLP (phylogeny inference package) v. 3.6*. Seattle, WA: Department of Genetics, University of Washington.
- Fitch, W. & Margoliash, E. 1967 Construction of phylogenetic trees. *Science* **155**, 279–284. (doi:10.1126/science.155.3760.279)
- Gascuel, O. 1997 BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.* **14**, 685–695.
- Geyer, C. 1991 Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proc. of the 23rd Symp. Interface, Seattle, WA*, pp. 156–163. Fairfax Station, VA: Interface Foundation of North America.
- Goldman, N. & Yang, Z. 1994 A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**, 725–736.
- Gonnet, G. H., Cohen, M. A. & Benner, S. A. 1992 Exhaustive matching of the entire protein sequence database. *Science* **256**, 1443–1445. (doi:10.1126/science.1604319)
- Gonnet, G. H., Hallett, M. T., Korostensky, C. & Bernardin, L. 2000 DARWIN v. 2.0: an interpreted computer language for the biosciences. *Bioinformatics* **16**, 101–103. (doi:10.1093/bioinformatics/16.2.101)

- Guindon, S. & Gascuel, O. 2002 Efficient biased estimation of evolutionary distances when substitution rates vary across sites. *Mol. Biol. Evol.* **19**, 534–543.
- Halpern, A. L. & Bruno, W. J. 1998 Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol. Biol. Evol.* **15**, 910–917.
- Hasegawa, M., Kishino, H. & Yano, T. 1985 Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160–174. (doi:10.1007/BF02101694)
- Hillis, D. M. 1996 Inferring complex phylogenies. *Nature* **383**, 130–131. (doi:10.1038/383130a0)
- Holder, M. T., Sukumaran, J. & Lewis, P. O. 2008 A justification for reporting majority-rule consensus tree in Bayesian phylogenetics. *Syst. Biol.* **57**, 814–821. (doi:10.1080/10635150802422308)
- Huelsenbeck, J. 1995 Performance of phylogenetic methods in simulation. *Syst. Biol.* **44**, 17–48. (doi:10.2307/2413481)
- Jukes, T. H. & Cantor, C. R. 1969 Evolution of protein molecules. In *Mammalian protein metabolism* (ed. H. N. Munro), pp. 21–32. New York, NY: Academic Press.
- Kimura, M. 1980 A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111–120. (doi:10.1007/BF01731581)
- Kishino, H., Thorne, J. L. & Bruno, W. J. 2001 Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Mol. Biol. Evol.* **18**, 352–361.
- Kosiol, C., Holmes, I. & Goldman, N. 2007 An empirical codon model for protein sequence evolution. *Mol. Biol. Evol.* **24**, 1464–1479. (doi:10.1093/molbev/msm064)
- Kuhner, M. K. & Felsenstein, J. 1994 A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* **11**, 459–468.
- Lanave, C., Preparata, G., Saccone, C. & Serio, G. 1984 A new method for calculating evolutionary substitution rates. *J. Mol. Evol.* **20**, 86–93. (doi:10.1007/BF02101990)
- Lartillot, N. & Philippe, H. 2004 A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **21**, 1095–1109. (doi:10.1093/molbev/msh112)
- Lartillot, N., Blanquart, S. & Lepage, T. 2004 PHYLOBAYES 2.3. See http://www.lirmm.fr/mab/article.php?id_article=329.
- Liò, P., Goldman, N., Thorne, J. L. & Jones, D. T. 1998 PASSML: combining evolutionary inference and protein secondary structure prediction. *Bioinformatics* **14**, 726–733. (doi:10.1093/bioinformatics/14.8.726)
- Martin, A. & Palumbi, S. 1993 Protein evolution in different cellular environments: cytochrome *b* in sharks and mammals. *Mol. Biol. Evol.* **10**, 873.
- Naylor, G. J. P. & Brown, W. M. 1998 *Amphioxus* mitochondrial DNA, chordate phylogeny, and the limits of inference based on comparisons of sequences. *Syst. Biol.* **47**, 61–76. (doi:10.1080/106351598261030)
- Powell, M. J. D. 1964 An efficient method for finding the minimum of a function of several variables without calculating derivatives. *Comput. J.* **7**, 155–162. (doi:10.1093/comjnl/7.2.155)
- Ren, F., Tanaka, H. & Yang, Z. 2005 An empirical examination of the utility of codon-substitution models in phylogeny reconstruction. *Syst. Biol.* **54**, 808–818. (doi:10.1080/10635150500354688)
- Robinson, D. F. & Foulds, L. R. 1981 Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131–147. (doi:10.1016/0025-5564(81)90043-2)
- Ronquist, F. R. & Huelsenbeck, J. P. 2001 MrBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 754–755.
- Rzhetsky, A. & Nei, M. 1992 Statistical properties of the ordinary least-squares, generalized least-squares, and minimum-evolution methods of phylogenetic inference. *J. Mol. Evol.* **35**, 367–375. (doi:10.1007/BF00161174)
- Saitou, N. & Nei, M. 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425.
- Schöniger, M. & von Haeseler, A. 1995 Performance of the maximum likelihood, neighbor joining, and maximum parsimony methods when sequence sites are not independent. *Syst. Biol.* **44**, 533–547. (doi:10.2307/2413659)
- Sullivan, J., Abdo, Z., Joyce, P. & Swofford, D. 2005 Evaluating the performance of a successive-approximations approach to parameter optimization in maximum-likelihood phylogeny estimation. *Mol. Biol. Evol.* **22**, 1386–1392. (doi:10.1093/molbev/msi129)
- Swofford, D. L. 2001 *PAUP*: phylogenetic analysis using parsimony (and other methods), version 4.0.b10*. Sunderland, MA: Sinauer.
- Takahashi, K. & Nei, M. 2000 Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used. *Mol. Biol. Evol.* **17**, 1251–1258.
- Tamura, K. & Nei, M. 1993 Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**, 512–526.
- Tavaré, S. 1986 Some probabilistic and statistical problems in the analysis of DNA sequences. In *Some mathematical questions in biology—DNA sequence analysis* (ed. R. M. Miura), Providence, RI: American Mathematical Society.
- Voelker, G. & Edwards, S. V. 1998 Can weighting improve bushy trees? Models of cytochrome *b* evolution and the molecular systematics of pipits and wagtails (Aves: Motacillidae). *Syst. Biol.* **47**, 589–603. (doi:10.1080/106351598260608)
- Whelan, S. & Goldman, N. 2001 A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **18**, 691–699.
- Yang, Z. 1994 Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39**, 306–314. (doi:10.1007/BF00160154)
- Yang, Z., Goldman, N. & Friday, A. 1995 Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. *Syst. Biol.* **11**, 316–324.
- Yang, Z., Nielsen, R., Goldman, N. & Pedersen, A. 2000 Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**, 431–449.
- Yule, G. U. 1925 A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. *Phil. Trans. R. Soc. B* **213**, 21–87. (doi:10.1098/rstb.1925.0002)
- Zwickl, D. J. 2006 Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. PhD thesis, University of Texas, Austin, TX.