Syst. Biol. 66(2):218–231, 2017

DOI:10.1093/sysbio/syw074
Advance Access publication September 14, 2016

Phylogenetic Tree Estimation With and Without Alignment: New Distance Methods and Benchmarking

MARCIN BOGUSZ AND SIMON WHELAN*

Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, 752 36 Uppsala, Sweden
**Correspondence to be sent to: Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, 752 36 Uppsala, Sweden; E-mail: simon.whelan@ebc.uu.se.*

Received 3 February 2016; reviews returned 12 June 2016; accepted 23 August 2016
Associate Editor: David Bryant

Abstract.—Phylogenetic tree inference is a critical component of many systematic and evolutionary studies. The majority of these studies are based on the two-step process of multiple sequence alignment followed by tree inference, despite persistent evidence that the alignment step can lead to biased results. Here we present a two-part study that first presents PaHMM-Tree, a novel neighbor joining-based method that estimates pairwise distances without assuming a single alignment. We then use simulations to benchmark its performance against a wide-range of other phylogenetic tree inference methods, including the first comparison of alignment-free distance-based methods against more conventional tree estimation methods. Our new method for calculating pairwise distances based on statistical alignment provides distance estimates that are as accurate as those obtained using standard methods based on the true alignment. Pairwise distance estimates based on the two-step process tend to be substantially less accurate. This improved performance carries through to tree inference, where PaHMM-Tree provides more accurate tree estimates than all of the pairwise distance methods assessed. For close to moderately divergent sequence data we find that the two-step methods using statistical inference, where information from all sequences is included in the estimation procedure, tend to perform better than PaHMM-Tree, particularly full statistical alignment, which simultaneously estimates both the tree and the alignment. For deep divergences we find the alignment step becomes so prone to error that our distance-based PaHMM-Tree outperforms all other methods of tree inference. Finally, we find that the accuracy of alignment-free methods tends to decline faster than standard two-step methods in the presence of alignment uncertainty, and identify no conditions where alignment-free methods are equal to or more accurate than standard phylogenetic methods even in the presence of substantial alignment error. [Alignment-free; distance-based phylogenetics; pair Hidden Markov Models; phylogenetic inference; statistical alignment.]

Inferring phylogenetic trees from molecular sequence data is a fundamental method used in evolutionary and systematic studies. The resulting tree may provide direct insight into the evolutionary relationships between individual species, or may reflect important aspects of the history of the sequences, such as gene duplication (Bowers et al. 2003) and incomplete lineage sorting (Maddison and Knowles 2006). The tree is also a critical component of other studies, where it is a nuisance parameter when inferring adaptive evolution (Yang 2006), studying the acquisition of new functions (Conant and Wolfe 2008) and the dating speciation events (Dos Reis et al. 2015). Phylogeny estimation is a difficult task since it requires distinguishing between vast numbers of potential evolutionary histories using only molecular data from the relatively small number of extant sequences at our disposal. Many tree inference methods have been proposed and the current state-of-the-art approach is to perform tree inference through a two-step process of multiple sequence alignment (MSA) followed by statistical tree inference (Felsenstein 1988). This method, although widely used, has well-known limitations.

The aim of the first step is to identify homologous characters between sequences and produce a heuristic estimate of those homologies in a MSA. The problems with this step arise from at least two sources (Chatzou et al. 2015). First, the most widely used MSA methods (MSAMs) cannot cope with statistical uncertainty and only return a single point estimate of the MSA with no indication of its reliability. Often there are very large numbers of MSAs with very similar scores, and there are only limited means for comparing them and no means of testing whether MSAs are significantly different from one another (Thompson et al. 1999). The second problem is that MSAMs try to reach a compromise between a variety of competing goals, including identifying homologous residues and residues that share the same structure or function in a protein. Accurate identification of structural similarity does not guarantee the shared ancestry of residues (Morrison et al. 2015).

The second step typically uses only a single fixed MSA and a probabilistic substitution model to estimate the tree that best fits the observed sequences, either through clustering based on pairwise distance estimates or through joint estimation of the tree and model parameters from all the sequences at once. Many substitution models have been developed, each capturing important aspects of the evolutionary process, such as rate variation between sites (Yang 1994), different rates of substitution between nucleotides (Hasegawa et al. 1985; Tavaré 1986), and the averaged substitution rates between amino acids (Whelan and Goldman 2001; Le and Gascuel 2008). The majority of research on sequence evolution has been done studying only this step under the strict assumption that the MSA is correct and all differences are down to substitutions in the sequence's history. Multiple studies have shown that uncertainty and inevitable error in MSA introduces bias at many levels, including tree estimates (Hossain et al. 2015), the accuracy of branch lengths (Blackburne and Whelan 2013), and the detection of adaptation using

dN/dS (Markova-Raina and Petrov 2011). The most popular approach to mitigating these problems is to try to remove uncertainly aligned regions using third-party filtering programs, such as Heads or Tails (Landan and Graur 2007) or GUIDANCE (Penn et al. 2010), but a recent study shows that filtering might actually lead to worse estimates (Tan et al. 2015). Other ways to alleviate the problem are integrating over some of the uncertainty in the MSA (Blackburne and Whelan 2013) and iteratively attempt to improve the MSA and the tree (Edgar 2004; Liu et al. 2011).

Several alternatives have been proposed to the computational convenience and speed of the two-step approach. The first comes from the early realization that MSA and phylogeny are the same problem (Sankoff and Kruskal 1983), which led to methods that combine alignment and tree estimation using models that capture insertions, deletions, and substitutions (Thorne et al. 1991, 1992). These models led to statistical alignment tools like BAli-Phy (Redelings and Suchard 2005) and StatAlign (Novák et al. 2008), which overcome the limitations of conditioning on a single MSA using Bayesian inference coupled with a sophisticated MCMC sampler to simultaneously estimate tree topologies, alignments, and model parameters. Although these tools account for statistical uncertainty in the alignment and the phylogeny, it comes at great computational expense with even relatively small-scale analyses taking days to run.

Another set of approaches have attempted to avoid assigning homology altogether, producing a set of so-called "alignment-free" methods. Instead these methods specify the distance between pairs of sequences based on simple similarity measures, and then use those pairwise distances to infer a tree topology. The similarity measures include concepts like the compression-based Lempel-Ziv (LZ) complexity (Otu and Sayood 2003) and an information theory-based average common substring (ACS) metric (Ulitsky et al. 2006), but the most popular is to calculate the relative occurrences of k-mers (Vinga and Almeida 2003). These similarity methods have been widely studied and have shown to be successful when estimating trees, mostly through simulation studies (Höhl and Ragan 2007). There are, however, no studies that systematically compare these alignment-free methods with more conventional two-step approaches to the tree estimation problem.

The aim of this study is 2-fold. First we present PaHMM-Tree (pairwise Hidden Markov Model Tree estimation, pronounced *palm-Tree*, available at http://paHMM-Tree.tk or http://marbogusz.github.io/paHMM-Tree/), a neighbor joining-based method that takes distances from pairwise statistical alignment to strike a compromise between the accuracy of full statistical alignment and the computational speed and ease of the distance-based methods. Second, we use a simulation approach to compare the accuracy of distance and tree estimation under PaHMM-Tree with a selected range of other phylogenetic methods, including standard two-step methods, statistical alignment, and alignment-free methods, which to the best of our knowledge is the first time all of these methods have been systematically compared. We begin by comparing the performance of PaHMM-Tree to methods that take a known "true" MSA from simulation. We find that pairwise distance estimates estimated using PaHMM-Tree on average have similar average accuracy and variance as distances estimated using standard maximum likelihood (ML) methods with the true alignment. Furthermore, the tree estimates from PaHMM-Tree compare favorably to those obtained from the two-step process on the known MSA: providing marginally more accurate estimates than trees estimated from ML estimates of pairwise distances, and worse estimates than full ML approaches using RAxML, a state-of-the-art tree inference tool (Stamatakis 2014). Next we examine the whole range of methods for the case when the MSA is not known. For closely related sequences, where the MSA is easy to estimate unambiguously, we find the two-step process tends to work well. For more divergent sequences, we find the performance of the two-step process declines more rapidly than other methods, leaving the statistical aligner BAli-Phy and PaHMM-Tree as the most accurate methods. Under all of the conditions in this study, the alignment-free methods perform worse than all of the other methods.

Methods

Computing the Likelihood of a Pair of Unaligned Sequences using a Pair-HMM

In order to infer the evolutionary distance between a pair of unaligned sequences x and y we require a probabilistic model and a method of statistical inference. Our model, summarized in Fig. 1, is based on the pair-HMM used in BAli-Phy and is most easily understood as a generative model. Pair-HMMs are approximate models that can be used to describe the evolution over time $t$from sequence x to sequence y using the match, insert, and delete "hidden" states, which capture the homology relationships between the pair of sequences. The evolutionary process is set up to assume time reversibility, so the probability of generating sequence y conditional on an initial sequence x is the same as generating sequence x conditional on the initial sequence y.

The match state generates (emits) a pair of homologous characters, represented as "++", and the standard phylogenetic substitution model within the match state can then be used to generate the substitution history of those characters. As usual, this substitution model contains a set of exchangeability parameters describing the rates of substitutions between characters, an equilibrium distribution of the characters obtained empirically from the observed sequences, an α parameter describing Γ-distributed rates across sites, and a time parameter $d_{x,y}$ which describes the evolutionary distance between the sequences in units of expected number of substitutions per site (Yang 2006).

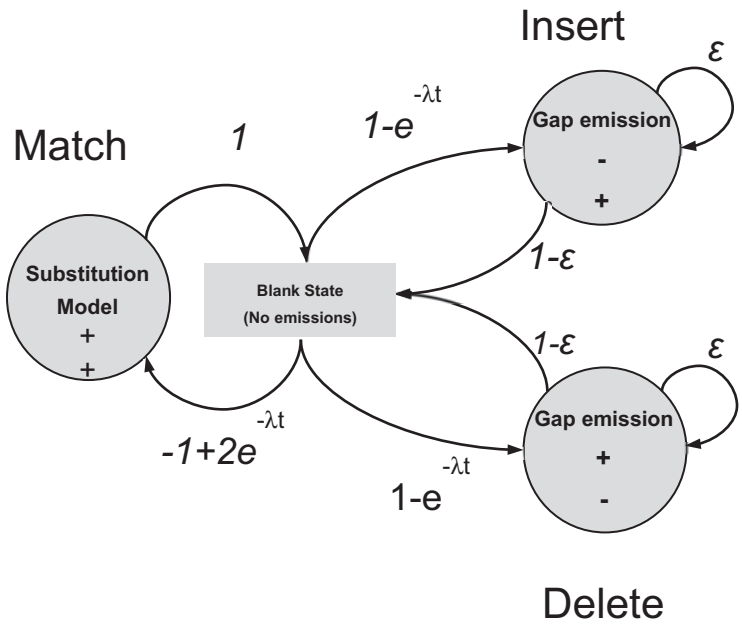

FIGURE 1. Model of substitutions, insertions, and deletions. Match state outputs characters in both sequences according to the substitution model. Insert state emits only a character in sequence 2 and delete state only emits a character in sequence 1. Evolutionary distance t expressed in expected number of substitutions per site. Insertion/deletion rate and length distribution represented by λ and ε parameters, respectively.

The insert and delete hidden states generate insertion and deletion events relative to the original sequence that are represented as the gain and loss of a set of characters from that sequence, respectively. The probability of transitioning into an insertion state represents the occurrence of an insertion directly after a site, which occurs at the rate $1-e^{-\lambda d_{x,y}}$, where λ captures the rate of insertion occurring relative to the substitutions and we place a limit on the maximum value of $d_{x,y}$ so $1-e^{-\lambda d_{x,y}}$ cannot exceed 0.5. This hard bound means that the compound of $\lambda d_{x,y}$ (the expected number of insertions and deletions for each substitution) cannot exceed 0.69, which seems reasonable since for these high values the sequences are effectively saturated with insertions and deletions. The ε parameter specifies the length of the insertion or deletion, which is geometrically distributed and independent of the evolutionary divergence. The inserted characters are drawn at random from the equilibrium distribution specified by the substitution model in the match state. Deletions occur in a similar manner, also with rate λ relative to the substitution rate, but instead remove an ε-based geometrically distributed length of sequence.

This formulation of the insertion and deletion process assumes that each character in a sequence can undergo at most a single insertion or deletion, meaning that there are no overlapping insertions or deletions. This assumption will not hold for very divergent sequences or sequences with relatively high values of λ, although we note our results suggest that this approximation works better than assuming a single fixed MSA.

In order to use this generative model for inference, we use the likelihood of a pair of sequences—taken as the probability of x and y being generated by the pair-HMM—and compute it using the Forward algorithm (Durbin et al. 1998). This likelihood represents the sum of the probabilities of all the paths through the pair-HMM and is equal to integration across all possible alignments between x and y. The likelihood conceptually breaks down into two components: the substitution likelihood and the insertion/deletion likelihood. The former is obtained directly from a standard substitution model in the match state, but also controls the frequency of characters in the root sequence and emissions from the insert state. The latter is a product of the transition probabilities between the hidden states, and controls the presence and length of gapped regions. In our implementation we allow a choice between two nucleotide substitution models, HKY85 and GTR (Hasegawa et al. 1985; Tavaré 1986), whereas for amino acids, we offer the LG model (Le and Gascuel 2008). Our software can be easily extended to allow for any time reversible substitution model. Pairwise divergence times and all other model parameters are obtained within the standard ML inference framework, using a set of heuristics described below to obtain faster estimates.

PaHMM-Tree: Implementation and Optimization

PaHMM-Tree is a command line program that takes a set of unaligned FASTA sequences as input and outputs a Newick tree along with a lower triangular matrix of pairwise distances in PHYLIP format. As with most statistical tree inference programs, the substitution model needs to be specified before the analysis starts. The user is also able to perform analysis using across site

rate heterogeneity, which uses a discrete Γ-distribution, and to specify the number of distinct Γ categories. PaHMM-Tree also allows for any combination of user-defined model parameters, including insertion/deletion rate and length distribution. Every state of our pair-HMM is represented by an m times n rectangular dynamic programming (DP) matrix, where m is the length of sequence x and n is the length of y. These matrices are used by HMM algorithms, whose computational complexity is proportional to the area of those matrices.

In order to obtain starting points in subsequent analyses we require an approximate starting distance for $d_{x,y}$, which we acquire from the fractional k-mer distances between the pair of sequences x and y:

$$k_{\text{dist}}(x,y) = 1 - \frac{\sum_{\text{k}_i \in \text{kmer}} \min(k_i(y), k_i(x))}{\min(|x|,|y|) - \text{kmer}_{\text{size}} - 1}, \quad (1)$$

where $k_i(x)$ is the number of times k-mer k_i occurs in sequence x and $|x|$ is the length of that sequence. In order to find a good choice of k-mer size we conducted extensive simulations, and found that 7-mers tended to work best for nucleotides, whereas 4-mers tended to performed well for amino acids. In order to obtain an evolutionary distance in terms of expected number of substitutions per site from $k_{\text{dist}}(x,y)$ we performed a range of simulations under known distances with evolutionary models with substitutions, insertions, and deletions, then fitted a curve to find the relationship between known evolutionary distances and the values of $k_{\text{dist}}(x,y)$.

During distance estimation we assume that the evolutionary process is homogeneous and stationary, meaning that the same model parameters can be used across all pairs of sequences. Jointly estimating the distances and the model parameters is possible, but computationally extremely slow so we take a heuristic approach where the model parameters are estimated from a subset of sequences and then taken as fixed when estimating the complete set of pairwise distances. In order to accurately estimate the Γ-distribution shape parameter, α, we pick a small set of triplets of sequences to perform substitution model analysis. (We note that Wu and Susko 2010, proved generic identifiability for the GTR+Γ model from a set of pairwise comparisons, but we in a practical setting that using only pairs requires long sequences to return even moderately accurate α estimates.) Our approach attempts to identify non-overlapping triplets of sequences with moderate branch lengths from our k-mer distances. We first run a number of Forward calculations for a small set of model parameter combinations: two predefined indel rate parameters λ, three alpha parameter values and three branch length modifiers for k-mer-based distance estimate. We choose the best combination of the aforementioned parameters that maximizes the forward likelihood across all pairs in our triplet set. For nucleotides we use the HKY85 model with κ parameter set to 2.0 and a fixed geometric gap length parameter of 0.5 for the gap length distribution. Based on these Forward calculations, we run the Backward algorithm and calculate posterior probabilities for every pair of residues in each of the three pairwise alignments per triplet. Finally, we use marginalized posterior decoding approach, as described in Lunter et al. (2008), to identify high confidence triplets of homologous sites across all pairs, which are then used to estimate the substitution model using the standard Felsenstein tree likelihood (Felsenstein 1981). To estimate the indel parameters, λ and ε, we only use the set of pairwise alignments from our triplets and maximize the likelihood of state transitions given the previously estimated divergence times on the triplets.

The next stage of estimating the full set of pairwise distances based on these fixed model estimates also involves some heuristic steps. To effectively calculate forward likelihoods we use a banding approach to limit the area of the DP matrix based on a Forward–Backward computation under reasonable starting conditions that identifies regions of high posterior probability alignments. Our initial bands are based on k-mer distance estimates and could potentially ignore some reasonable alignments under different divergences, but checks under real and simulated data suggest this problem does not occur. The final bands are constructed column-wise by finding the cell of highest posterior probability and greedily adding cells above and below in the column until a predefined value of a posterior probability is reached. Using this approach, the bandwidth is not constant and depends on the posterior probability distribution.

Our tree building approach uses the standard BioNJ algorithm (Gascuel 1997), which offers the best tree accuracy among the available neighbor joining algorithms—see Supplementary Fig. S1, available on Dryad at http://dx.doi.org/10.5061/dryad.n5r49, for an accuracy comparison. We obtain the entries in the matrix of evolutionary distances in a straightforward manner from the previously optimized divergence times. Assuming that sequences x and y have similar lengths n, the computational complexity of a single pairwise distance estimation is $O(n^2)$ and use of narrow bands reduces the complexity to $O(n \times c)$, where c is a constant relating to the width of the band. In a data set of $l$sequences, the number of possible pairs is $(l-1) \times l$, however, since the likelihood of a pair x,y is equal to the likelihood of y,x we only need $(l-1) \times l \times 0.5$ pairwise calculations with the quadratic computational complexity.

Other Programs and Methods Examined

Our study uses a set of three popular MSAMs: the progressive aligner MAFFT version 7.2 running on its most accurate settings—L-INS-i (Katoh and Standley 2013); the evolutionary aligner PRANK v.140603 under default settings (Löytynoja and Goldman 2008); and the consistency T-coffee version 11.00.8cbe486 under default settings (Notredame et al. 2000).

The ML tree analysis for the two-step tree inference approach was performed using RAxML version 8.2 (Stamatakis 2014) with the discrete Γ across site rate heterogeneity enabled (GTRGAMMA and PROTGAMMALG parameters). In case of pairwise distance estimation from sequence alignments, we used numerical optimization within the ML framework using GTR model for nucleotides and LG for amino acids.

For distance-based inference from aligned sequences, we used the BioNJ algorithm implemented in PhyD* software package (Criscuolo and Gascuel 2008). From our tests, BioNJ tends to provide slightly better tree estimates to the standard NJ algorithm (Supplementary Fig. S1, available on Dryad).

Statistical joint alignment and tree estimation was performed using BAli-Phy version 2.3.7 with the RS07 insertion/deletion model (Redelings and Suchard 2007) and a discrete Γ substitution rate distribution (–smodel=GTR+gamma[4] parameter). SATé, as a representative of iterative alignment and tree refinement algorithm, was run using default settings using MAFFT for the alignment step and FastTree for phylogenetic inference (Price et al. 2009).

For alignment-free methods we used a custom Python implementation of k-mer counts with the distance metric being the same as in Equation (1). LZ and ACS methods were computed using decaf+py software package (Höhl et al. 2006) and tree estimates were performed using PhyD* and the BioNJ algorithm.

Simulation Conditions and Accuracy Measures

In order to study the accuracy of tree inference methods we simulate data using INDELible version 1.03 (Fletcher and Yang 2009) under the general time reversible model with Γ-distributed across sites rate heterogeneity (GTR+Γ) for nucleotides and LG+Γ for amino acids. The nucleotide model parameters are inspired by mammalian genomes with GTR exchangeability and nucleotide frequency parameters coming from Arbiza et al. (2011) and discrete Γ-shape parameters similar to those from Goldman and Whelan (2002). Indel lengths are modeled using negative binomial distribution and both the rate and length parameters are set to reflect the values from Taylor et al. (2004). To create some variety in the data properties examined, all parameters are drawn independently for every replicate from normal distributions with a standard deviation of 10% of the parameter value, except for the Γ-shape parameter where we allow a 20% standard deviation. Unrooted phylogenetic trees are generated using a Yule pure birth process. To ensure these trees have the desired height we transformed the birth parameter to the expected tree height and draw trees until a sample is within 10% of the desired height. The accuracy of inferred trees is measured using the Robinson–Foulds distance (Robinson and Foulds 1981) scaled to a range between 0 and 1 to allow for comparisons between trees of different sizes, with a measure of 1 representing identical trees. Tree distances based on branch lengths were not considered since alignment-free distances are not in units of expected number of substitutions per site and the branches from NJ-based trees may not reflect those from full statistical inference.

Results

Pairwise Distance Estimation under PaHMM-Tree

First we examine the performance of PaHMM-Tree when estimating pairwise distances, a critical step when inferring distance-based trees using clustering methods, and compare performance to other state-of-the-art methods. We simulate pairs of nucleotide and amino acid sequences with the expected number of substitutions per site ranging from 0.1 to 2.9 in 0.2 increments, with 50 independent simulation samples per increment. Figure 2 shows the ML distance estimates using the correct model parameters obtained using PaHMM-Tree, ML distances estimated from the true pairwise alignment, and ML distances estimated from the MAFFT, PRANK, and T-COFFEE pairwise alignments. This set of comparisons is intended to represent all possible comparable approaches to pairwise distance inference when the MSA is known and not known. For both sets of sequence data the true MSA, as expected, provides accurate distance estimates, although the pair-HMM distances are of a similar accuracy. For amino acid data, our approach returns more accurate estimates in around 60% of the cases compared with the true alignment, whereas for nucleotides, the performance is nearly identical, with around 50% data sets examined having more accurate estimates under the pair-HMM. Moreover, the variances of the pairwise distance estimates from our pair-HMM implementation and from the true alignment-based two-step analysis are similar (see Supplementary Fig. S2, available on Dryad), suggesting that any increase in variance attributable to not knowing the alignment is off-set by the increase in information obtained from using insertion and deletion information. In order to assess the impact of insertion/deletion information, Supplementary Fig. S2, available on Dryad also shows the variance of distance estimates obtained from our pair-HMM model using the likelihood of the true alignment including both the substitution and indel processes. This method has the lowest variance, which is noticeably lower in the case of nucleotide data, suggesting that insertions and deletions provide useful information for the distance inference.

The results from the alignment-based methods tend to depend on what type of sequence data are used. For nucleotides, MAFFT led to overestimates of the distance, possibly due to over-alignment, whereas PRANK and T-COFFEE both led to substantial underestimates. The alignments and distances from amino acid sequences are easier to estimate than those from nucleotide sequences due to the larger state-space and because the probability

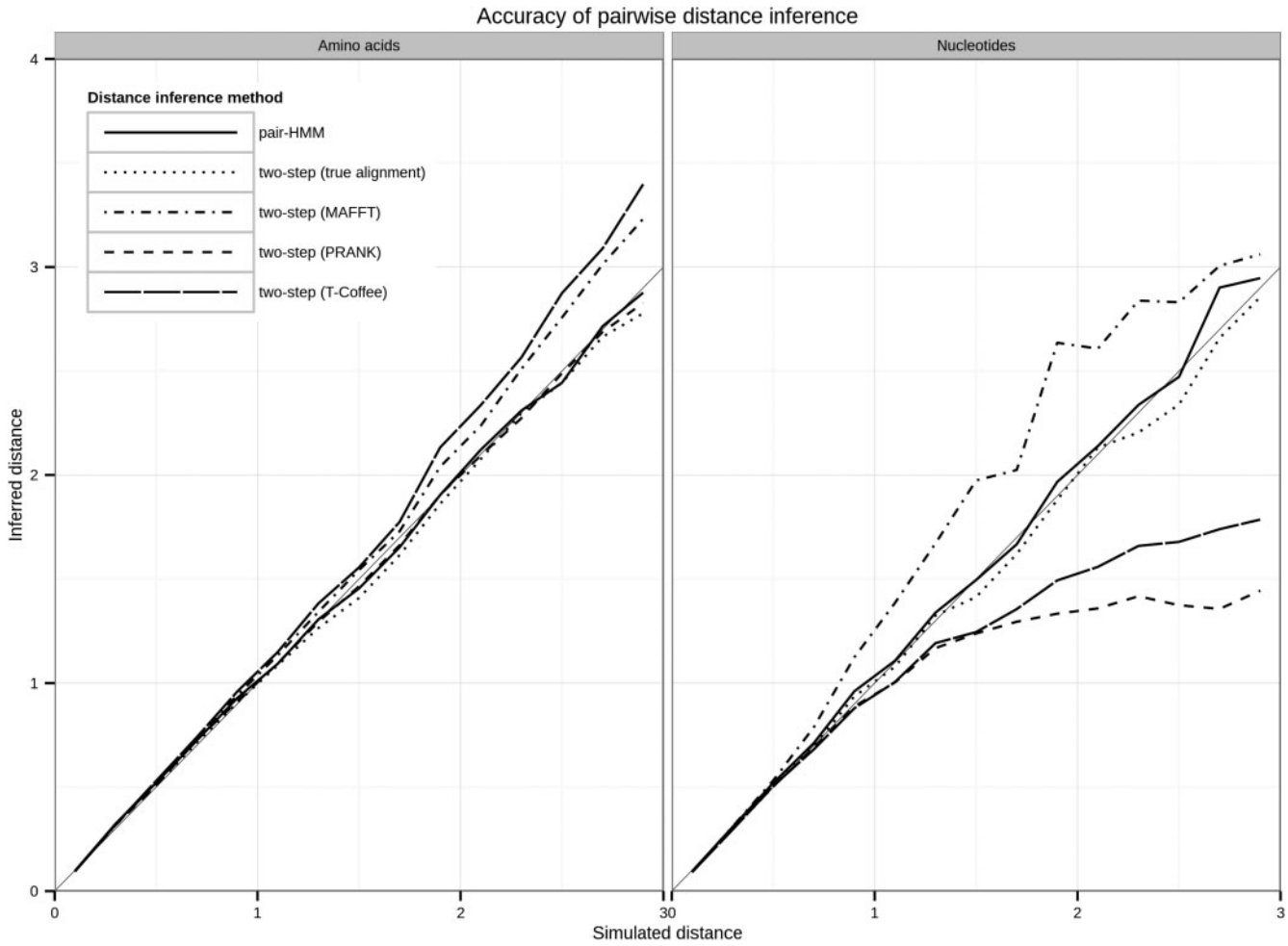

FIGURE 2. Median pairwise distance estimates for 500 character-long sequences with 50 simulation replicates per distance category. Distances inferred using GTR model using true simulated substitution model parameters within the ML framework. Amino acid data simulated and inferred under LG model. Distances in expected number of substitutions/site.

of back mutation is lower. PRANK provides good estimates of the distances even for very divergent sequences, whereas use of T-COFFEE and MAFFT result in moderate over-estimates for divergent sequences. The improvement of PaHMM-tree over MSA-based inference under high divergent scenarios may be attributable to its lack of alignment error (false positive and false negative homologies) The additional indel information also seems to play a role in the recovery of accurate distances. From these results we conclude that PaHMM-Tree can obtain accurate distance estimates—on a par with using the true alignment—even for very divergent sequences, and that the use of fixed alignments tend to result in less accurate distance estimates.

Inferring Phylogenies from Known MSAs

In order to distinguish between the accuracy of MSAMs and the performance of tree inference, we need a baseline measure where the sequence alignment is known. Figure 3 (and Supplementary Fig. S3, available on Dryad) shows the relative performance of different tree inference methods based on the true MSA produced from simulations based on 16 (and 64) sequences. These sequence numbers are intended to represent the case of medium-small (and medium-large) phylogenies, with the tree heights ranging from 0.25 to 16 expected substitutions per site. This reflects the range of very easy to extremely difficult MSA and tree inference problems.

As expected and widely discussed elsewhere (Whelan et al. 2001; Felsenstein 2004), the joint statistical likelihood approach of RAxML performs best when the true MSA is known for both amino acid and nucleotide sequences. A Bayesian approach would be expected to perform similarly. The performance of the BioNJ distance methods based on the known MSA and PaHHM-Tree is similar, reflecting the results of the previous section that show that distance estimates are similarly accurate with and without the MSA. We also find that the increased accuracy of distances inferred from amino acid is also reflected in tree estimates,

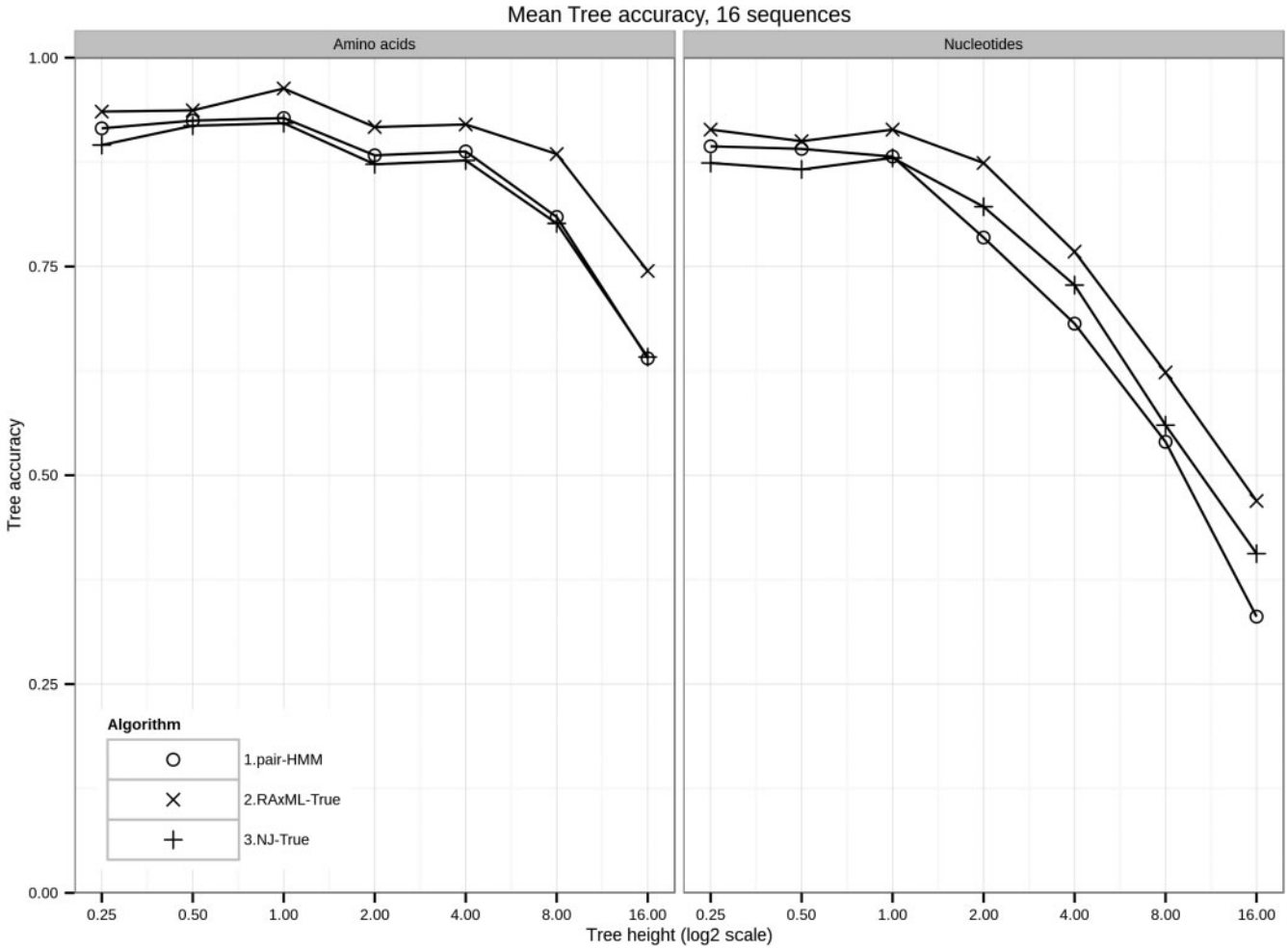

FIGURE 3. Average accuracy for 16 sequence trees estimated from true simulated alignments and model parameters with 50 independent tree simulations per tree height category. PaHMM-Tree uses raw sequence data and true simulated model parameters.

with all methods providing more accurate tree estimates when using amino acid sequences. For neighbor joining, the 64 sequence set has a marginally lower accuracy to its smaller counterpart for both nucleotides and amino acids, which can be attributed to greater variance in pairwise distance estimates relative to shorter internal branches (Gascuel and Steel 2006) and no increase in the information for each pair of sequences. For RAxML the accuracy results for bigger trees are, in contrast, better, reflecting the increased information from jointly considering all the sequences on the true alignment matrix (Supplementary Fig. S3, available on Dryad). From these results and the previous section we conclude that PaHMM-Tree performs similarly to distance-based methods where the true alignment is known, but statistical methods that consider all the sequences at once, such as ML, will tend to be superior to both.

Inferring Phylogenies and MSAs

In order to examine the performance of PaHMM-Tree and the two-step methods of tree inference in a scenario reflective of real-world data analysis, we benchmarked the accuracy of tree inference when the sequence alignment is not known. The group of methods we examine are based on the MSA and tree inference steps that we found to be most accurate in a preliminary analysis. For MSA we examine MAFFT, PRANK, and T-COFFEE, as representatives of progressive, evolutionary, and consensus MSAMs, respectively. For tree inference we examine PaHMM-Tree, BioNJ based on distances estimated from the inferred MSA, and ML inference, conducted using RAxML on the inferred MSA. Figure 4 (and Supplementary Fig. S4a, available on Dryad) shows the accuracy of these different methods over the same tree heights as the previous section.

As expected, the performance of two-step methods deteriorates substantially when both steps are performed, particularly for more divergent sequences where MSA and tree inference is more difficult. The relative accuracy of the methods does not seem to depend on the size of the data set. In both 16 and 64 sequence trees (Fig. 4 and Supplementary Fig. S4a, available on Dryad, respectively), the performance

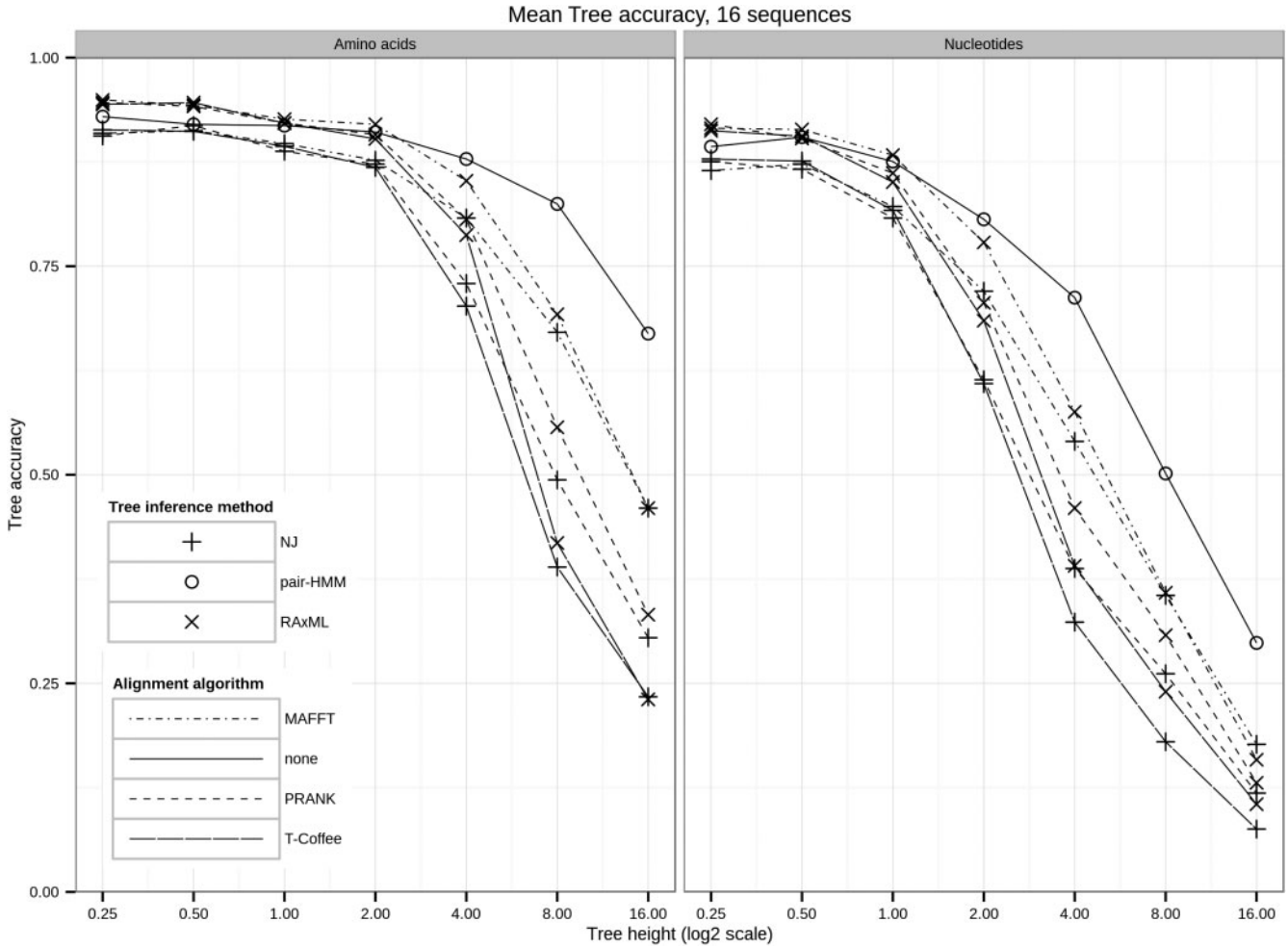

FIGURE 4. Average accuracy for 16 sequence trees estimated from three different types of MSAs coupled with RAxML and BioNJ. PaHMM-Tree uses raw sequence data and estimates all the model parameters. Sample size of 50 replicates per tree height category.

of PaHMM-Tree falls between NJ and RAxML for short trees and outperforms the other methods for highly divergent phylogenies. We also investigated the performance of Fast Statistical Alignment version 1.15.9 (Bradley et al. 2009), which is a MSA method based on pair-HMMs and sequence annealing but it's performance was similar to MAFFT for low divergences, but decayed faster for higher divergences (Supplementary Material, Figs. S4b and c, available on Dryad).

The best of the two-step methods investigated here tends to be MAFFT paired with RAxML, which is the best performing method on relatively shallow trees up to around 1.0 for nucleotide and 2.0 for amino acid sequences. After this point the performance of all methods drops off substantially, but PaHMM-Tree's falls at a slower rate and it becomes the most accurate tree inference method for more divergent sequences. It is clear from comparing the results in this and the previous section that the methodological choice of the MSA step is more important for tree accuracy than the tree inference methodology for the most divergent sequences, again emphasizing that the accuracy of tree inference is critically dependent on the quality of the MSA. From these results we conclude that the MSA followed by ML tree inference works very well for closely related sequences, but as sequences become more divergent PaHMM-Tree's ability to cope with MSA uncertainty becomes more important than the specific method of tree inference.

Comparisons between Full Statistical Alignment and Alignment-Free Methods

Statistical alignment has long held the promise of more accurate tree estimation than standard approaches to phylogenetic inference. Figure 5 shows the accuracy of the full statistical aligner BAli-Phy, using both consensus trees and maximum a posteriori (MAP) estimates. The figure also includes the performance of SATé, which iteratively improves the MSA and the tree estimate. We compare estimates obtained under these methods to PaHMM-Tree and the best of the two-step methods (MAFFT-RAxML). Computing power

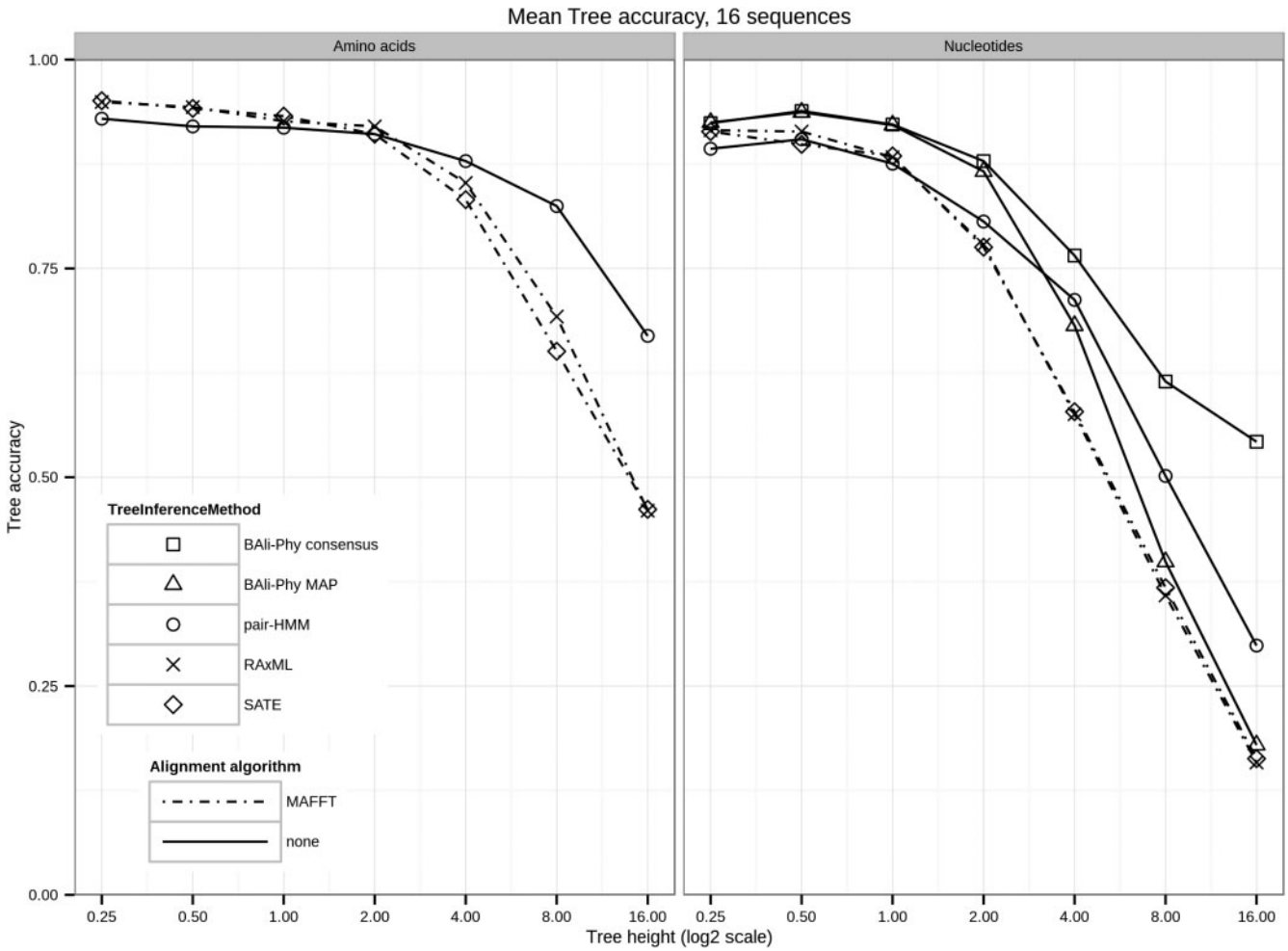

FIGURE 5. Average accuracy for 16 sequence trees estimated using best two-step process, PaHMM-Tree, Bayesian joint statistical alignment and tree estimation, and SATé using default settings. Sample size of 50 replicates per tree height category.

limitations mean that BAli-Phy could only be run on 16-sequence nucleotide data, and attempts to run on the same size amino acid data were prohibitively computationally expensive, with each single run taking upwards of 50 h to burn-in and obtain accurate estimates.

The results for the nucleotide data confirm that the BAli-Phy tends to provide the most accurate tree estimates for low and moderate divergent sequences, reflecting BAli-Phy's efficient use of joint information about insertions, deletions, and substitutions from the whole set of sequences when inferring trees. PaHMM-tree, on the other hand, considers only pairs of sequences independently, which limits the amount of information available for distance estimates. As the sequences become more divergent this relative accuracy declines, particularly for MAP tree estimates, with PaHMM-Tree providing more accurate estimates for very divergent sequences. This loss of performance may be due to the priors in BAli-Phy expecting more closely related sequences. Although we cannot reject it, the decline in performance does not seem related to poor mixing since rerunning the MCMC or letting run for much longer results in the same estimates.

Our results also show that BAli-Phy consensus trees tend to be equally or more accurate than MAP trees. This accuracy occurs because the consensus trees tend to be mostly star topologies due to the wide variety of tree estimates in the posterior, whereas the MAP estimates are fully resolved. Our measure of tree accuracy, the normalized Robinson–Foulds distance from the true tree, reflects both the number of bipartitions present in an inferred tree but not the reference tree (false positives) and the number of bipartitions found in the reference tree but not the comparison tree (false negatives). For bifurcating trees the number of false positives and negatives are equal, but star topologies have no false negatives and their accuracy cannot be lower than 0.5.

The accuracy of SATé is mostly indistinguishable from using MAFFT-RAxML for both small and large data sets (Fig. 5 and Supplementary Fig. S5, available on Dryad respectively) and not comparable to the accuracy

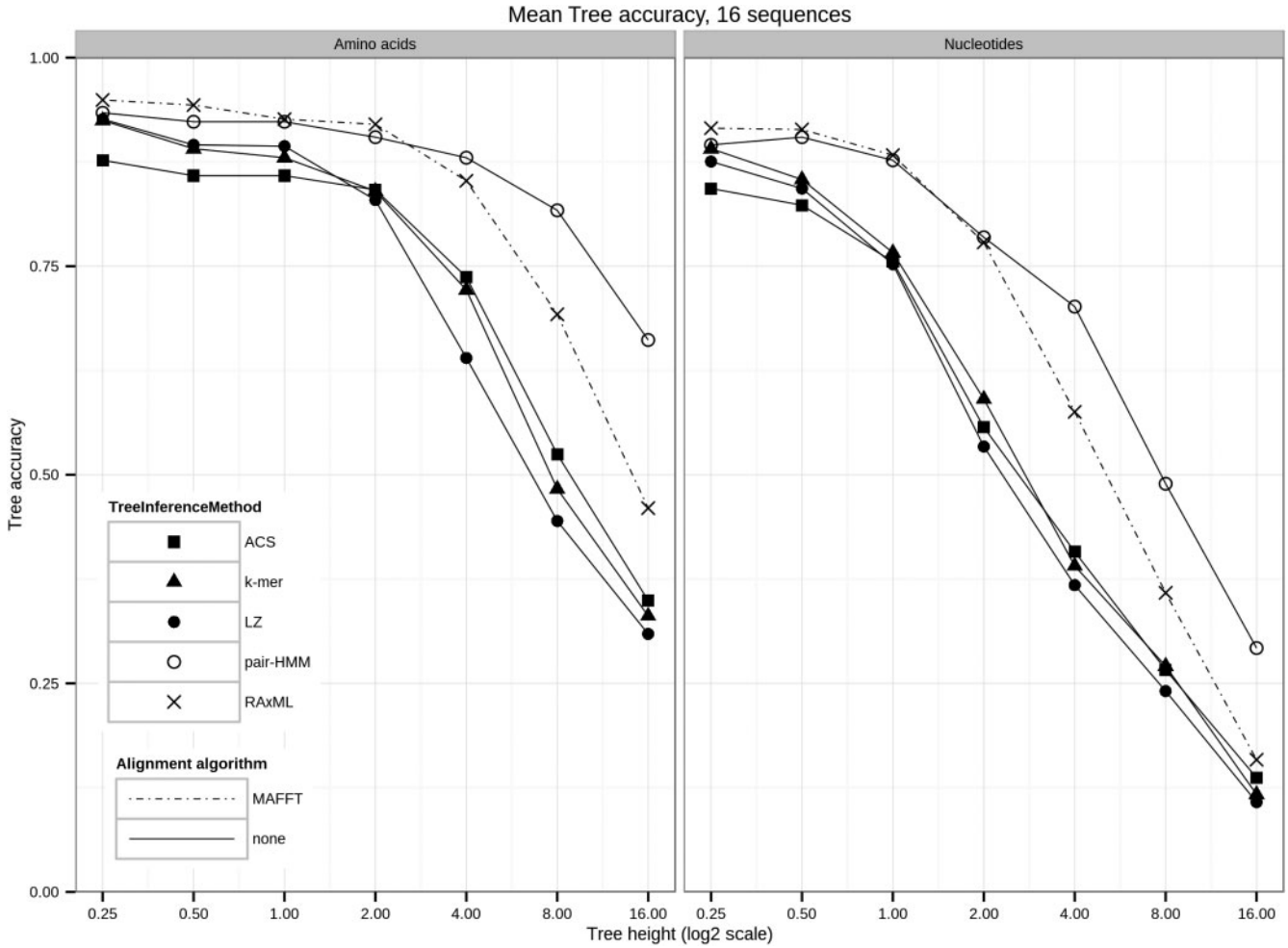

FIGURE 6. Average accuracy for 16 sequence trees estimated using best two-step process, PaHMM-Tree and 3 various alignment-free methods. Sample size of 50 replicates per tree height category.

achieved by the full statistical alignment program BAli-Phy. This result may be expected since SATé does not attempt to use insertion and deletion information, nor integrate across MSA uncertainty, so lacks many of the advantages of BAli-Phy. Moreover, SATé's authors suggest in their documentation that the method is designed to perform better on very large data sets, rather than the size of data sets we examine here.

There have also been many published studies demonstrating that alignment-free methods (ACS; LZ complexity; and k-mer counts) are suitable for inferring phylogenies, although few attempts to compare these methods to more traditional phylogenetic methods have been made. In Fig. 6 (and Supplementary Fig. S6, available on Dryad), we compare the accuracy of PaHMM-Tree and MAFFT-RAxML to a range of these alignment-free methods.

All of the alignment-free methods perform relatively similarly and in all cases are less accurate than both the two-step process and paHMM-Tree. Even for deep trees where the MSA is difficult to estimate and the two-step methods perform relatively poorly, it still appears preferable to use MAFFT-RAxML over existing standard alignment-free methods. PaHMM-Tree is more accurate than alignment-free methods over all divergences.

Computational Performance of Methods

Speed is a major consideration when estimating phylogenies, so we measured the execution times of paHMM-Tree and compared them with those of RAxML and other methods. All analyses were run in single core mode on an AMD Opteron 6220 CPU-based machine. RAxML is known for its speed due to an elaborate set of heuristics and, in the case of nucleotides, was superior in both data sets with our method being approximately 1.5–3 times slower depending on sequence divergence. For amino acids the relationship differs: RAxML can take as long as 15 h for an analysis of a 64-taxa tree compared with 10 min of paHMM-Tree CPU execution time (Fig. 7).

The performance of PaHMM-Tree is correlated with the tree height, where highly divergent trees take more time. This may be attributed to the fact that initial distance estimates from k-mers are more reliable for

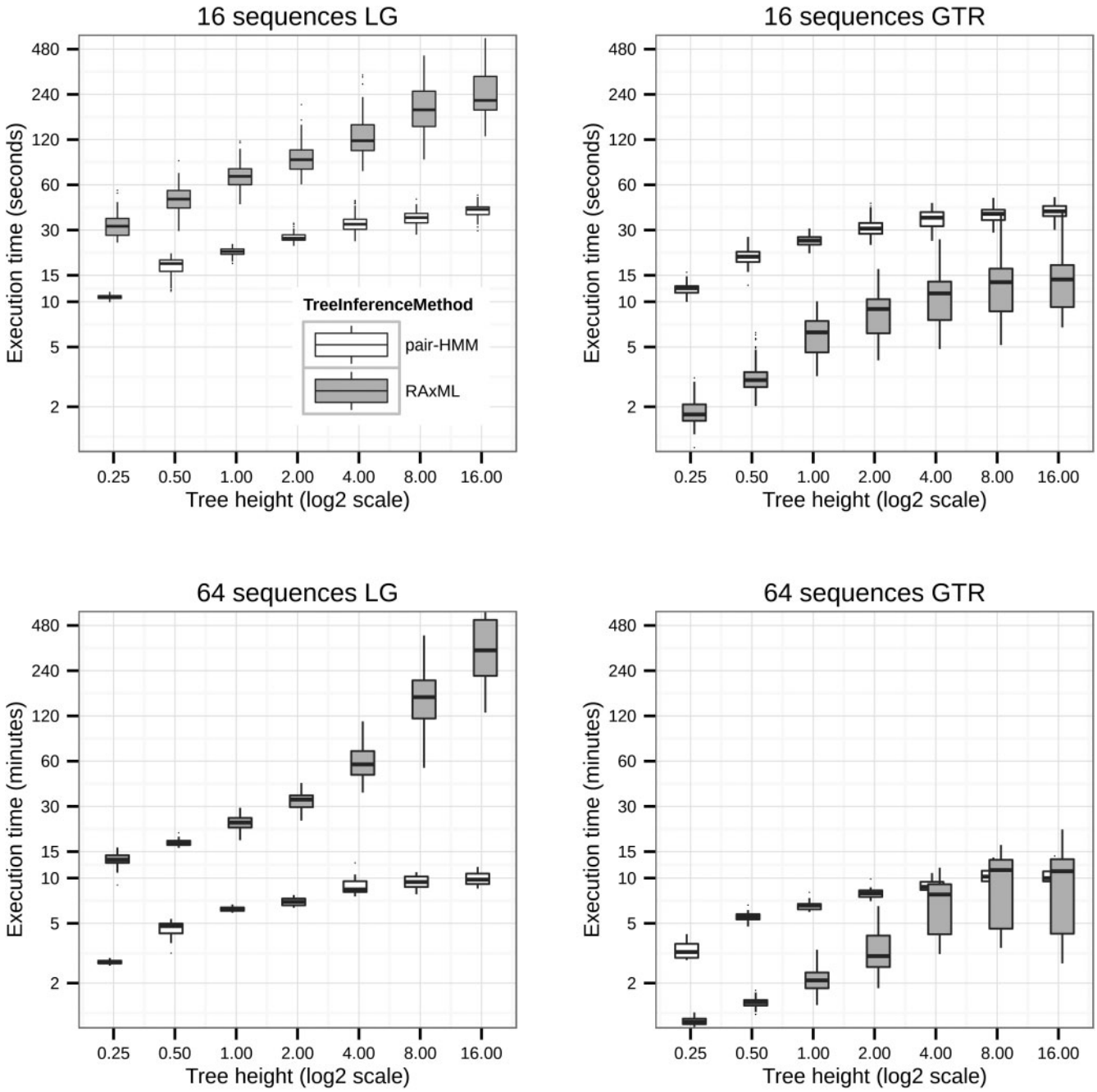

FIGURE 7. Execution times for RAxML and paHMM-Tree.

conserved trees, as well as flatter forward likelihood surfaces for long branches. Also with highly divergent pairs of sequences, the bands on the DP matrix need to be wider to accommodate more paths, further reducing the speed of each forward likelihood calculation. Table 1 shows the average execution times for the methods used in this study. It is clear that PaHMM-Tree provides fast tree estimates, with no difference in execution times between amino acid and nucleotide data. Bayesian statistical alignment is very time consuming even for a relatively small 16-sequence data set and unfeasible for larger trees. Alignment-free methods have good computational performance, with the k-mer distance implementation offering the best performance.

DISCUSSION

The interaction between MSA and the accuracy of phylogenetic inference continues to be a major source of bias and uncertainty in phylogenetic and phylogenomic studies (Wong et al. 2008; Hossain et al.

TABLE 1. Average execution time in minutes in cumulative processor core time

Method	16 Sequences		64 Sequences	
	NT	AA	NT	AA
paHMM-Tree	0.48	0.45	7.59	6.96
MAFFT + BioNJ	0.04	0.04	0.05	0.5
MAFFT + RAxML	0.22	1.72	5.59	65.68
ACS + BioNJ	0.03	0.03	0.5	0.5
LZ + BioNJ	0.2	0.3	3.31	5.24
k-mer + BioNJ	0.01	0.01	0.02	0.02
BAli-Phy	7740	N/A	N/A	N/A
SATé (FastTree)	0.16	0.18	2.42	1.91

Notes: Mean CPU time includes tree inference process and alignment/distance estimation step where applicable. Values averaged across the whole data sets of 16 and 64 sequence trees for nucleotide (NT) and amino acid (AA) data.

2015). Here we present a new method that infers pairwise distances between sequences by integrating across all possible alignments using an explicit evolutionary model. We chose a relatively simple pair-HMM model previously implemented in BAli-Phy to make use of well established HMM algorithms and so our results are directly comparable to BAli-Phy. Simulations show that the distances our method obtains are very accurate and seem unbiased, even for very divergent sequences where alignment is very hard. Moreover, this accuracy is obtained at a minimal cost to the variance of those estimates, with the additional distance information from insertions and deletions seeming to offset the uncertainty introduced by not knowing the alignment.

The improved accuracy in pairwise distance estimates carries through to tree inference, with the trees calculated from our pair-HMM inferred distance matrix being more accurate than standard pairwise methods the overwhelming majority of the time, especially at larger evolutionary distances. For closely related sequences, this improvement may be attributable to incorporating indel information, whereas for more distantly related sequences the incorporation of alignment uncertainty is also very important. These findings strongly suggest there is little reason to perform distance-based analyses on aligned sequences unless computational considerations make it impractical. Our timings suggest this may occur in the mid- to high-hundreds of sequences or for longer sequences.

There is a more complex relationship between the performance of PaHMM-Tree and joint statistical methods, such as the two-step process with RAxML tree inference and statistical alignment using BAli-Phy. Our analyses comparing PaHMM-Tree with RAxML applied to the true alignment show that jointly estimating the tree from all sequences is beneficial, unambiguously supporting the orthodoxy that joint ML and Bayesian inference should outperform distance-based methods (Felsenstein 2004; Yang 2006).

When the MSA is unknown the performance of different methods is highly dependent on the sequence divergence. For closely related sequences, the alignment step is relatively easy and joint statistical methods perform well, providing a clear improvement over all distance-based methods. In the case of nucleotide sequences where we have the computational resources to run the statistical aligner BAli-Phy, we find it performs better than the two-step process with RAxML; further suggesting there is significant information available for tree inference in insertions and deletions even for closely related sequences. The relative performance of PaHMM-Tree improves as sequence divergence increases, surpassing that of the two-step process for moderately-to-highly diverged sequences where standard MSAMs and downstream tree estimates suffer from excessive false positive and false negative homologies. For more extreme levels of divergence, our results also suggest PaHMM-Tree outperforms full statistical alignment, although this may be due to the inadequacy of the Bayesian priors for such extreme distances, the poor mixing of the MCMC chains, or both.

Our final benchmarks compare a range of alignment-free methods to more traditional phylogenetic methods. Although there has been extensive interest in alignment-free methods over several years (Vinga and Almeida 2003; Haubold 2014), we find no evidence that they perform better than standard phylogenetic methods in any of the conditions we examine. The accuracy of alignment-free methods, both in terms of distance estimation (shown in Supplementary Fig. S7, available on Dryad) and tree inference, tends to be marginally worse than standard pairwise phylogenetic methods for closely related sequences, but the performance of alignment-free methods declines more rapidly than that of other methods. This decline may be attributable to alignment-free methods working on measures of similarity, which count the observed changes between sequences (p-distances), rather than evolutionary distance, which corrects for "multiple-hits" and measures the number of substitutions. The evolutionary distances estimated from statistical methods account for these multiple substitutions in their evolutionary model, but it is difficult to imagine a valid evolutionary model that translates to observed changes in k-mer similarity and the level of sequence compression. As a consequence there may be little prospect of an equivalent multiple-hits correction for alignment-free methods. These limitations do not mean that alignment-free methods do not have a useful role to play. They can be fast and can be applied readily to large numbers of sequences, so are already widely used for guide-tree construction by MSA methods (Katoh et al. 2002; Edgar 2004). Their ability to handle very long sequences also means they may prove a suitable tool for comparing closely related genomes. Our results suggest, however, there is little reason to use alignment-free methods for tree inference when more standard phylogenetic and phylogenomic studies are applicable.

For a benchmarking study, such as this one, it is also important to discuss the limitations of the experimental design, both in terms of the methods tested and the

conditions examined, and how they affect the generality of the conclusions drawn. The first, and possibly most important, limitation is the simulation methodology with respect to insertions and deletions. INDELible, in common with all other phylogenetic simulators of its type, assumes that the insertion and deletion process is uniform along the sequence. This assumption can be relaxed in INDELible by specifying blocks of sequence with different model parameters, but this approach misses the types of epistatic interactions that lead to the indel patterns observed in real proteins. A coil region in a protein, for example, is more likely to have fixed insertions and deletions than an α-helix, but the length distribution and pattern of these changes is still constrained by range of lengths acceptable for that coil and how specific residues in that coil interact with the rest of the protein. Similarly these epistatic interactions tend to limit the number of residues or nucleotides occurring at constrained sites, resulting in distant sequences have a higher percentage identity than would be expected from a general model, such as WAG or LG (Whelan and Goldman 2001; Le and Gascuel 2008). The lack of block like patterns of conservation and wider distribution of characters at individual sites in simulated data may make aligning simulated data substantially harder than real data, suggesting that the performance of the two-step methodology with RAxML may be better than our results suggest. This effect is unlikely to affect the relative performance of two-step distance-based NJ versus PaHMM-Tree since both methods would benefit from the easier alignment and PaHMM-Tree continues to be able to exploit the indel events in addition to substitutions.

A second limitation is that our study only examines relatively small numbers of sequences compared with the many thousands or tens of thousands that can be considered in some evolutionary and functional studies (Smith et al. 2011; Chatzou et al. 2015). The transition from 16 to 64 sequences provides some insight into the relative behavior of the methods, but the transition to very large numbers of sequences is a qualitatively different problem due to the computational constraints it imposes and how different MSA methods and phylogenetic methods address those problems. Our previous work has shown that some MSAMs fail to cope with large numbers of sequence (Hossain et al. 2015), so the relative performance of MSA-based methods to alignment-free methods could potentially change. It seems unlikely that full statistical alignment methods will ever be able to analyze such large numbers of sequences in a reasonable time frame. At present PaHMM-Tree cannot cope with such large numbers of sequences, but further improvements to its heuristics could improve matters.

A final, limitation of our PaHMM-Tree method is that it provides no fast way of estimating the support values for the internal branches of the tree. The dependence between sites induced by insertions and deletions means we cannot use standard nonparametric bootstrapping since it relies on sampling independent columns with replacement. Parametric bootstrapping, where new samples are drawn from the model, may be possible, but its applicability will be subject to the limitations of simulation methodology discussed above and computational constraints. We also note that this problem also occurs for alignment-free methods, which suffer the added detriment that they cannot use parametric bootstrapping because they do not use an explicit probabilistic model.

Despite these limitations, PaHMM-Tree provides a unique approach to the phylogenetic tree inference problem, providing fast and accurate tree inference based on an explicit phylogenetic model without conditioning on an MSA. Although at present these trees do not have support measures associated with them, they offer valuable preliminary data analysis and could offer a new way of estimating guide trees for difficult phylogenetic problems. We hope that future improvements to PaHMM-Tree will help to alleviate some of its computational limitations, for example by implementing the mBed algorithm to reduce the number of distance calls, allowing it to process larger data sets (Blackshields et al. 2010), or through the implementation of anchor points using suffix trees in the pairHMMs, allowing it to work with longer sequences more quickly (Gusfield 1997).

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.n5r49.

References

Arbiza L., Patricio M., Dopazo H., Posada D. 2011. Genome-wide heterogeneity of nucleotide substitution model fit. Genome Biol. Evol. 3:896–908.

Blackburne B.P., Whelan S. 2013. Class of multiple sequence alignment algorithm affects genomic analysis. Mol. Biol. Evol. 30:642–653.

Blackshields G., Sievers F., Shi W., Wilm A., Higgins D.G. 2010. Sequence embedding for fast construction of guide trees for multiple sequence alignment. Algorithms Mol. Biol. 5:21.

Bowers J.E., Chapman B.A., Rong J., Paterson A.H. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature 422:433–438.

Bradley R.K., Roberts A., Smoot M., Juvekar S., Do J., Dewey C., Holmes I., Pachter L. 2009. Fast statistical alignment. PLoS Comput. Biol. 5:e1000392.

Chatzou M., Magis C., Chang J.-M., Kemena C., Bussotti G., Erb I., Notredame C. 2015. Multiple sequence alignment modeling: methods and applications. Brief. Bioinform. bbv099.

Conant G.C., Wolfe K.H. 2008. Turning a hobby into a job: how duplicated genes find new functions. Nat. Rev. Genet. 9:938–950.

Criscuolo A., Gascuel O. 2008. Fast NJ-like algorithms to deal with incomplete distance matrices. BMC Bioinformatics 9:166.

Durbin R., Eddy S.R., Krogh A., Mitchison G. 1998. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge, UK: Cambridge university press.

Edgar R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368–376.

Felsenstein J. 1988. Phylogenies from molecular sequences: inference and reliability. Annu. Rev. Genet. 22:521–565.

Felsenstein J. 2004. Inferring phylogenies. Sunderland: Sinauer Associates.
Fletcher W., Yang Z. 2009. INDELible: a flexible simulator of biological sequence evolution. Mol. Biol. Evol. 26:1879–1888.
Gascuel O., Steel M. 2006. Neighbor-joining revealed. Mol. Biol. Evol. 23:1997–2000.
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.
Goldman N., Whelan S. 2002. A novel use of equilibrium frequencies in models of sequence evolution. Mol. Biol. Evol. 19:1821–1831.
Gusfield D. 1997. Algorithms on strings, trees and sequences: computer science and computational biology. Cambridge, UK: Cambridge University Press.
Hasegawa M., Kishino H., Yano T. 1985. Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160–174.
Haubold B. 2014. Alignment-free phylogenetics and population genetics. Brief. Bioinform. 15:407–418.
Höhl M., Ragan M.A. 2007. Is multiple-sequence alignment required for accurate inference of phylogeny? Syst. Biol. 56:206–221.
Höhl M., Rigoutsos I., Ragan M.A. 2006. Pattern-based phylogenetic distance estimation and tree reconstruction. Evol. Bioinform. Online 2:359.
Hossain a S.M.M., Blackburne B.P., Shah A., Whelan S. 2015. Evidence of statistical inconsistency of phylogenetic methods in the presence of multiple sequence alignment uncertainty. Genome Biol. Evol. 7:2102–2116.
Katoh K., Misawa K., Kuma K., Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30:3059–3066.
Katoh K., Standley D.M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30:772–780.
Landan G., Graur D. 2007. Heads or tails: a simple reliability check for multiple sequence alignments. Mol. Biol. Evol. 24:1380–1383.
Le S.Q., Gascuel O. 2008. An improved general amino acid replacement matrix. Mol. Biol. Evol. 25:1307–1320.
Liu K., Warnow T.J., Holder M.T., Nelesen S.M., Yu J., Stamatakis A.P., Linder C.R. 2011. SATe-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. Syst. Biol. 61:90–106.
Löytynoja A., Goldman N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science 320:1632–1635.
Lunter G., Rocco A., Mimouni N., Heger A., Caldeira A., Hein J. 2008. Uncertainty in homology inferences?: assessing and improving genomic sequence alignment. Genome Res. 18:298–309.
Maddison W.P., Knowles L.L. 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55:21–30.
Markova-Raina P., Petrov D. 2011. High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 *Drosophila* genomes. Genome Res. 21:863–874.
Morrison D.A., Morgan M.J., Kelchner S.A. 2015. Molecular homology and multiple-sequence alignment: an analysis of concepts and practice. Aust. Syst. Bot. 28:46.
Notredame C., Higgins D.G., Heringa J. 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 302:205–217.
Novák Á., Miklós I., Lyngsø R., Hein J. 2008. StatAlign: an extendable software package for joint Bayesian estimation of alignments and evolutionary trees. Bioinformatics 24:2403–2404.
Otu H.H., Sayood K. 2003. A new sequence distance measure for phylogenetic tree construction. Bioinformatics 19:2122–2130.
Penn O., Privman E., Landan G., Graur D., Pupko T. 2010. An alignment confidence score capturing robustness to guide tree uncertainty. Mol. Biol. Evol. 27:1759–1767.
Price M.N., Dehal P.S., Arkin A.P. 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol. Biol. Evol. 26:1641–1650.
Redelings B.D., Suchard M.A. 2005. Joint Bayesian estimation of alignment and phylogeny. Syst. Biol. 54:401–418.
Redelings B.D., Suchard M.A. 2007. Incorporating indel information into phylogeny estimation for rapidly emerging pathogens. BMC Evol. Biol. 7:40.
Dos Reis M., Donoghue P.C.J., Yang Z. 2015. Bayesian molecular clock dating of species divergences in the genomics era. Nat. Rev. Genet. 17:71–80.
Robinson D.F., Foulds L.R. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.
Sankoff D., Kruskal J.B. 1983. Time warps, string edits, and macromolecules: the theory and practice of sequence comparison. Addison-Wesley Publishing Co.: Reading, Massachusetts.
Smith S.A., Beaulieu J.M., Stamatakis A., Donoghue M.J. 2011. Understanding angiosperm diversification using small and large phylogenetic trees. Am. J. Bot. 98:404–414.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.
Tan G., Muffato M., Ledergerber C., Herrero J., Goldman N., Gil M., Dessimoz C. 2015. Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference. Syst. Biol. 64:1–33.
Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Lect. Math. Life Sci. 17:57–86.
Taylor M.S., Ponting C.P., Copley R.R. 2004. Occurrence and consequences of coding sequence insertions and deletions in mammalian genomes. Genome Res. 14:555–566.
Thompson J., Plewniak F., Poch O. 1999. BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. Bioinformatics 15:87–88.
Thorne J.L., Kishino H., Felsenstein J. 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. J. Mol. Evol. 33:114–124.
Thorne J.L., Kishino H., Felsenstein J. 1992. Inching toward reality: an improved likelihood model of sequence evolution. J. Mol. Evol. 34:3–16.
Ulitsky I., Burstein D., Tuller T., Chor B. 2006. The average common substring approach to phylogenomic reconstruction. J. Comput. Biol. 13:336–350.
Vinga S., Almeida J. 2003. Alignment-free sequence comparison—a review. Bioinformatics 19:513–523.
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.
Whelan S., Liò P., Goldman N. 2001. Molecular phylogenetics: state-of-the-art methods for looking into the past. Trends Genet. 17: 262–272.
Wong K.M., Suchard M.A., Huelsenbeck J.P. 2008. Alignment uncertainty and genomic analysis. Science 319:473–476.
Wu J., Susko E. 2010. Rate-variation need not defeat phylogenetic inference through pairwise sequence comparisons. J. Theor. Biol. 263:587–589.
Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39:306–314.
Yang Z. 2006. Computational molecular evolution. Oxford: Oxford University Press.