Syst. Biol. 0(0):1–19, 2017

DOI:10.1093/sysbio/syx068

Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation

Huai-Chun Wang[1,2,3], Bui Quang Minh[4], Edward Susko[1,3], and Andrew J. Roger[2,3,*]

[1]*Department of Mathematics and Statistics, 6316 Coburg Road;* [2]*Department of Biochemistry and Molecular Biology, 5850 College Street, Dalhousie University, Halifax, Nova Scotia B3H 4R2, Canada;* [3]*Centre for Comparative Genomics and Evolutionary Bioinformatics, Dalhousie University, Halifax, Nova Scotia B3H 4R2, Canada; and* [4]*Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories, University of Vienna and Medical University of Vienna, Austria*
Huai-Chun Wang and Bui Quang Minh contributed equally to this article.
[*]*Correspondence to be sent to: Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia B3H 4R2, Canada; E-mail: andrew.roger@dal.ca.*

Received 21 October 2016; reviews returned 9 July 2017; accepted 2 August 2017
Associate Editor: Mark Holder

Abstract.—Proteins have distinct structural and functional constraints at different sites that lead to site-specific preferences for particular amino acid residues as the sequences evolve. Heterogeneity in the amino acid substitution process between sites is not modeled by commonly used empirical amino acid exchange matrices. Such model misspecification can lead to artefacts in phylogenetic estimation such as long-branch attraction. Although sophisticated site-heterogeneous mixture models have been developed to address this problem in both Bayesian and maximum likelihood (ML) frameworks, their formidable computational time and memory usage severely limits their use in large phylogenomic analyses. Here we propose a posterior mean site frequency (PMSF) method as a rapid and efficient approximation to full empirical profile mixture models for ML analysis. The PMSF approach assigns a conditional mean amino acid frequency profile to each site calculated based on a mixture model fitted to the data using a preliminary guide tree. These PMSF profiles can then be used for in-depth tree-searching in place of the full mixture model. Compared with widely used empirical mixture models with k classes, our implementation of PMSF in IQ-TREE (http://www.iqtree.org) speeds up the computation by approximately $k/1.5$-fold and requires a small fraction of the RAM. Furthermore, this speedup allows, for the first time, full nonparametric bootstrap analyses to be conducted under complex site-heterogeneous models on large concatenated data matrices. Our simulations and empirical data analyses demonstrate that PMSF can effectively ameliorate long-branch attraction artefacts. In some empirical and simulation settings PMSF provided more accurate estimates of phylogenies than the mixture models from which they derive. [Long-branch attraction; long-branch repulsion; maximum likelihood; mixture model; posterior mean site frequency; site heterogeneity.]

Analyzing large numbers of orthologous genes from many taxa is necessary to resolve deep phylogenetic problems in the tree of life including the origins of major groups such as animals (Pisani et al. 2015), plants (Wickett et al. 2014), fungi (Kuramae et al. 2006), the intra-/inter-domain relationships of eukaryotes (Brown et al. 2013), Archaea (Raymann et al. 2015), and Bacteria (Daubin et al. 2002). Phylogenomic approaches can drastically reduce stochastic errors associated with small data sets used in traditional phylogenetic studies and have led to substantial advances in our knowledge of the tree of life (Delsuc et al. 2005; Philippe et al. 2011). Unfortunately, merely increasing the number of sequences in the analysis is sometimes insufficient to resolve many difficult phylogenetic questions (Philippe et al. 2011). In addition to sequence alignment quality control, proper taxon and gene sampling, and outgroup choice, it is crucial to use models that can faithfully capture the underlying amino acid substitution process in phylogenomic analyses.

The substitution process at an individual alignment position (site) in a protein is often mathematically modeled as a stochastic Markov process with an empirically-defined single rate matrix of amino acid exchangeabilities (e.g., the JTT matrix of Jones et al. 1992, the WAG matrix of Whelan and Goldman 2001 or LG matrix described in Le and Gascuel 2008) and the stationary frequencies of the amino acids in the data. The process at different sites is traditionally considered as homogeneous whereby the same process operates over all sites with the exception of variable evolutionary rates amongst sites which is modeled by a gamma distribution discretized into a few rate categories of equal probability. While site-homogeneous models are computationally efficient and have been widely used for phylogenetic inference, they are not biologically realistic. Different sites in proteins have different structural or functional constraints that result in different preference for specific amino acids at sites. Some sites can be occupied by almost any residue, while others appear to be restricted to a limited subset of amino acids or just one particular residue (Halpern and Bruno 1998; Lartillot and Philippe 2004; Lartillot et al. 2007; Wang et al. 2008). Such site-specific amino acid preferences are not captured by the standard empirical rate matrices. Indeed it was shown that the number of amino acid states at a given site in real data sets is, on average, smaller and the frequencies of these states are less uniform than those expected under JTT+F+Γ (Lartillot et al. 2007; Wang et al. 2008). Halpern and Bruno (1998) showed that site-homogeneous models tend to underestimate long distances between the sequences more than short distances when the data sets are simulated under site-specific frequency profiles. In other words, site-homogeneous models are sensitive to mutational saturation of the sequences whereby it is difficult to distinguish true phylogenetic signal from homoplasy (i.e., multiple independent origins of the same character state at a homologous position in different lineages). Such sensitivity tends to lead to long-branch attraction (LBA) artifacts in phylogenetic estimation (Lartillot et al. 2007; Wang et al. 2008).

To address these issues, partitioned (Yang 1996; Pupko et al. 2002; Lanfear et al. 2012) and mixture models (Lartillot and Philippe 2004; Le et al. 2008a, 2008b; Wang et al. 2008) have been developed to account for the substitution heterogeneity across genes and sites respectively. A widely used site-heterogeneous model is CAT, a Bayesian mixture model which assumes each site has its own set of equilibrium amino acid frequencies (Lartillot and Philippe 2004). The equilibrium frequency vectors are independently and identically drawn from an unknown distribution, which is nonparametrically estimated using a Dirichlet process model. The CAT model has been implemented in Phylobayes (Lartillot et al. 2013) in a Markov chain Monte Carlo (MCMC) framework to allow joint estimation of the mixture components and the exchange rate matrix for a given data set. The resulting CAT+GTR model has been shown to not only fit data better than the site-homogeneous models but also alleviate LBA bias (Lartillot et al. 2007). The model has since been widely used for phylogenomic reconstruction and helped resolve some long-standing phylogenetic questions (e.g., Struck et al. 2011). However, a common concern with this approach is that the MCMC chains have convergence difficulties for large data sets (Kocot et al. 2011; Pisani et al. 2015; Whelan et al. 2015; Whelan and Halanych 2016), which limits its application.

Models accounting for site heterogeneity in maximum likelihood framework include a mixture of the substitution rate matrices predefined for different secondary structures and surface accessibility (Goldman et al. 1998; Le et al. 2008b; Le and Gascuel 2010), or for different site rates (Le et al. 2012) and a mixture of amino acid site frequency profiles (Wang et al. 2008, 2014; Le et al. 2008a). The latter approaches have been implemented in QmmRAxML (Wang et al. 2008), PhyML (Guindon et al. 2010) and IQ-TREE (Nguyen et al. 2015). To ease the computational burden, fixed empirical amino acid frequency vectors are used rather than being estimated from the data. Several such empirical frequency profiles (cF4, C10, C20 to C60) have been proposed (Le et al. 2008a, 2008b; Wang et al. 2008, 2014). Both simulation and empirical studies have shown that the profile mixture models are more robust against the LBA bias than the single profile model that uses the overall data set frequencies or the equilibrium frequencies of the amino acid exchange rate matrix (Wang et al. 2008). However, relative to an analysis with a single frequency vector, RAM usage and computational time are effectively multiplied by k, where k is the number of components in the mixture. Except for the mixtures of structural matrices (e.g., EX2 and EX_EHO) or mixtures of four matrices for different site rates (such as LG4M and LG4X) that are relatively fast (Le and Gascuel 2010; Le et al. 2012), the mixture models with larger numbers of classes can become intractable to perform complete standard nonparametric bootstrap analyses, including tree searching, for very large phylogenomic data sets even when parallel or multicore implementations are used. In this study we propose several new models to approximate the profile mixture models and show that they are approximately $k/1.5$ faster than the profile mixture models and reduce the memory use by a factor of nearly k (when the number of taxa is larger than 30).

Materials and Methods

Modeling Site Heterogeneity: Posterior Mean Site Frequency and Posterior Maximum Site Frequency

A natural estimate of the frequency vector for a site is given by the observed frequencies of amino acids. However, such an estimate has several shortcomings. First, no adjustment is made for rate variation or phylogenetic relatedness of the taxa. For low rate sites, this leads to some amino acids occurring at inflated frequencies and others to have lowered frequencies relative to the true stationary distribution. Furthermore, unless the number of taxa is large and due to a positive dependence in amino acid presence for taxa related on a tree, there is a bias towards frequency profiles with a number of zero frequencies. Finally, experiences in simpler settings with numbers of parameters that increase as sample size increases (Neyman and Scott 1948) suggests that completely separate frequency estimates at each site may give rise to poor statistical properties. Indeed, Rodrigue (2013) noted such problems in obtaining reasonable parameter estimates when using ML to estimate site-specific amino acid frequencies in mutation-selection models.

In order to ensure that frequency estimates are nonzero and to avoid some of the other difficulties due to sparseness alluded to above, we introduce posterior mean site frequency (PMSF) which is computed under a mixture model given a guide tree. This effectively smooths the frequencies, making them more homogeneous, borrowing information from other sites rather than using sparse information from the site at hand. Specifically, the posterior probability for component j ($1 \leqslant j \leqslant k$) in the mixture is computed as

$$P(j|x) = \frac{w_j \times P(x|j)}{\sum_j w_j \times P(x|j)}, \quad (1)$$

where the w_j are the weights of the mixture component j and $P(x|j)$ is the probability of site pattern x under component j. The PMSF for site pattern x is defined by:

$$f_a(x) = \sum_j f_{aj} \times P(j|x), \quad (2)$$

where a indexes 20 amino acids and f_{aj} is the frequency for amino acid a in the mixture component j. While there are a finite number of f_{aj}, the $P(j|x)$ will vary over different x, giving different PMSFs for different sites.

For brevity we sometimes refer to PMSF as a model much like one would refer to LG+F as a model. More precisely, PMSF refers to both a method of estimation and a model that allows frequencies of amino acids to vary over sites and not necessarily in a manner consistent with a finite mixture. Indeed, an advantage with the PMSF is that it allows continuous variation of frequency vectors over sites. While mixtures with a finite number of classes are necessary for computational reasons, the uniqueness of the structural and functional constraints of sites in proteins suggests frequency vectors are better modeled as a continuous variation. Posterior means will tend to show continuous variation across sites; such behavior has been shown for posterior mean rate estimates calculated under a finite mixture by Susko et al. (2003).

An alternative approach considered in examples replaces (2) with posterior mode frequencies; we refer to such estimates as MAX estimates. For a give site pattern x, $f_a(x)$ is set to $f_{aj(x)}$, where $j(x)$ is the mixture component having the largest posterior probability $P(j|x)$. By contrast with PMSF, which assigns distinct frequency vectors to each site having a distinct site pattern, the number of unique MAX frequency vectors assigned to sites is at most the number of components k, independent of the alignment length. This leads to computational advantages compared with PMSF because, for any given edge, at most $k\times r$ substitution matrices need to be calculated where r is the number of site rate categories in the discrete gamma model.

Computational Efficiency of PMSF

Savings in runtime.—Computational savings arise for PMSF by comparison with a mixture model because rather than averaging the site-likelihoods over k frequency vectors as in a mixture model, the likelihood computation at a site uses one single frequency vector for each site likelihood calculation. On the other hand, PMSF has an additional start-up cost by comparison with single matrix models or mixtures that is due to obtaining eigen-decompositions of the rate matrices. The $O(c^3)$ cost of these decompositions must be repeated at each site where $c=20$ is the number of character states. Once this calculation is complete, however, it need not be repeated as additional trees are considered. Since the bulk of the computation in the course of considering multiple trees is devoted to likelihood evaluations, we compare computational cost of likelihood evaluation, assuming eigen-decompositions of the rate matrices. Regardless of the calculation—LG+F+Γ, PMSF or mixture—the total computational cost of likelihood evaluation is $n\times r$ times the cost of likelihood evaluation at a site and for a given rate category, where n is the number of unique site patterns and r the number of rate categories. Thus we need only compare the relative costs for a given site and rate category. The relative computational costs are calculated in the appendix and give that LG+F+Γ is expected to be at most 1.5 times as fast as PMSF and PMSF is expected to be at least $k/1.5$ times as fast as a k-component mixture. While it may seem that PMSF should give a cost comparable to LG+F+Γ, some exponentiation and multiplication operations need to be repeated over sites for PMSF but not for LG+F+Γ.

Savings in RAM usage.—The single matrix model requires $n\times r\times c\times(m-2)$ RAM to store the conditional likelihood vectors, where m is the number of taxa. The PMSF approach needs to additionally store the rate matrices, eigenvectors and inverse eigenvectors for all sites, which amounts to $3\times c^2\times n$. Therefore, compared with single matrix model the RAM usage under PMSF is multiplied by the following factor:

$$\left(1+\frac{3c}{r\times(m-2)}\right). \tag{3}$$

As m increases, the factor in (3) converges to 1, indicating the PMSF will use the same amount of the RAM as a single rate matrix model.

Note that while there are substantial savings in both RAM and runtime for tree-searching under PMSF, the full mixture models are required to derive these models in the first place by fitting on a "guide tree" thereby requiring the extra runtime and attendant RAM resources for the full mixture for this step (although see the memory saving technique described below). Nevertheless, for larger data sets this fitting phase is relatively quick compared with the tree-searching and bootstrapping phases of the analyses.

Software Implementation and Bootstrap Analysis

We have implemented the site-specific PMSF and MAX models in the phylogenetic inference program IQ-TREE (Nguyen et al. 2015), freely available at http://www.iqtree.org. To speed up likelihood computations, IQ-TREE employs vector operations and OpenMP (Open Multi-Processing), which parallelizes the computations over CPU cores. The PMSF or MAX models are executed by specifying a profile mixture model and a guide tree. The mixture model will first be fit on the guide tree and then the PMSF or MAX frequency profiles will be calculated. These site-specific profiles will then be used for site-likelihood calculations in an ML tree search using the standard methods available in IQ-TREE. Alternatively, one can use the −fs option in IQ-TREE to input their desired site frequency profiles instead of the guide tree, so that other approaches of obtaining site frequencies can be tested and trees are inferred.

To reduce the RAM requirement for the first fitting step of the mixture model, IQ-TREE implemented a memory saving technique (Izquierdo-Carrasco et al. 2012), where the minimal RAM consumption for likelihood computations is proportional to $\log_2(m)$

TABLE 1. Notation of PMSF and MAX models

Model	Guide tree	Study cases
PMSF0; MAX0	Known wrong tree corresponding to the LBA tree or the LBR tree depending on the setting	In simulation
PMSF1; MAX1	Known true tree	In simulation
PMSF2	ML tree estimated under LG+F+Γ	In simulation and empirical cases
PMSF3	ML tree estimated under LG+C20+F+Γ	In simulation and empirical cases

instead of m. This increases the computation time for the fitting step, but the extra time is small compared with the subsequent tree search. Moreover, IQ-TREE automatically adjusts the memory usage depending on the available computer RAM.

As the PMSF and MAX models have substantial computational savings in runtime, it is possible to conduct the standard nonparametric bootstrap analysis under the new models; this is infeasible for full profile mixture model with many components. To further ease the computational cost for bootstrap analysis under PMSF (or MAX), instead of refitting a full mixture model (e.g., LG+C20+F+Γ) to compute the PMSF profile for every bootstrap replicate, the site patterns and their associated PMSF frequency vectors were resampled from the original alignments. The resampled PMSF profiles are then applied to compute the bootstrap trees for the corresponding bootstrap replicate. Experiments showed that a full refitting of PMSF for each bootstrap replicate yielded nearly identical results to this "resampled PMSF strategy".

PMSF Model Notation

The PMSF posterior site frequency vectors are derived from a full profile mixture under a predefined guide tree. In the following simulation and empirical studies we will usually fit the LG+C20+F+Γ mixture to obtain the PMSF profiles. In a few cases the JTT+C20+F+Γ or LG+C60+F+Γ mixtures are used but such exceptions will be specified in the text. Based on the guide tree, we assign a digit number to the PMSF model to simplify the model notations (Table 1).

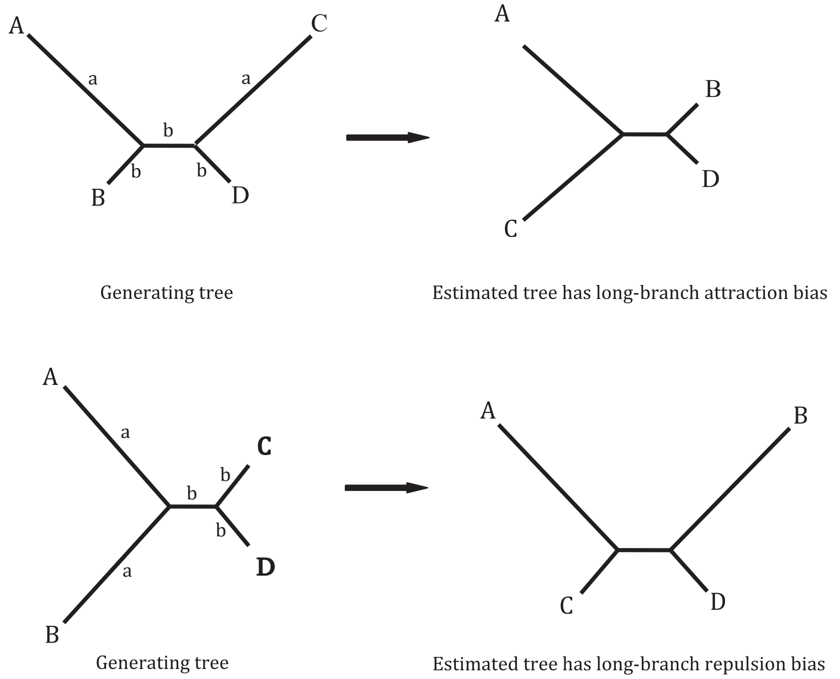
Simulated Data

Four taxa simulation under LG+C20+F+Γ.—One of the most commonly reported difficulties in phylogenetic estimation is LBA bias (Felsenstein 1978). To evaluate relative performance of the methods in LBA settings, we simulated protein alignments from four taxon trees having two long external branches separated (Fig. 1 upper panel—LBA simulations). Since methods that perform well in LBA settings might do so as a consequence of a bias towards trees with long branches apart (i.e., a long branch repulsion [LBR] bias; Susko et al. 2004), we also simulated data from trees of 4 taxa having two long external branches together (Fig. 1 lower panel—LBR simulations). The long branches varied in length from 0.1 to 1.0 with an increment of 0.1 and the short branches varied from 0.01 to 0.1 with an increment of 0.01. The internal branch has the same length as the two short external ones. For both LBA and LBR cases the sequences were simulated under LG+Γ4 (4 discrete gamma rates and alpha = 0.75) with a modified version of Seq-gen (Rambaut and Grassly 1997) that was adapted to generate sequence alignments under site-specific frequency profiles (SiteSpecific.seq-gen; http://www.mathstat.dal.ca/~hcwang/Procov/). The site-specific simulations were based on a 21-component mixture with 20 components from the C20 frequency classes (Le and Gascuel 2008) plus an additional class for the stationary amino acid frequencies of the LG matrix (LG+C20+F+Γ; Wang et al. 2014). Two sequence lengths were simulated for each data set: a short one with 1050 alignment sites (50 sites per frequency class) and a long one with 21,000 sites (1000 sites per frequency class). For each case, 100 data sets were generated for each pair of the long branch length (a) and short branch length (b) setting.

We then used the LG+F+Γ, LG+C20+F+Γ, and PMSF models separately to estimate the ML trees with IQ-TREE. To investigate the impact of using both good and poor guide trees to fit the mixture model initially for the PMSF model, we considered results when the guide tree was always the correct tree (PMSF1) and results when the guide tree was always a wrong tree (PMSF0). These two guide trees represent the best and worst scenarios that a PMSF model will be based on, although it is usually not possible to use the "correct tree" as guide tree for empirical data analysis. Therefore, we also used the ML tree estimated under the simpler LG+F+Γ as the guide tree (PMSF2) to evaluate the performance of PMSF on the simulated data. The results are summarized with heat maps.

Each cell of the heat map gives the accuracy of the estimation (i.e., the proportion of times the correct tree was estimated out of 100 simulations) when the corresponding a (y-axis) and b (x-axis) values were the branch lengths of the generating trees.

Four taxa simulation under LG+F+Γ.—For the two generating trees shown in Figure 1 we also simulated sequences under LG+F+Γ (with alpha = 0.75) to study the impact of overfitting, i.e., when the data are analyzed under the more complex LG+C20+F+Γ mixture and PMSF models. 100 short (1000 sites) and 100 long (20,000



Generating tree

Estimated tree has long-branch attraction bias

Generating tree

Estimated tree has long-branch repulsion bias

FIGURE 1. Four taxon generating trees used for simulation to induce long-branch attraction (LBA: upper panel) or long-branch repulsion (LBR: lower panel).

sites) alignments were simulated for each pair of branch lengths *a* and *b*.

Four taxon simulation under LG+C60+F+Γ.—Since real phylogenomic data are expected to have a larger range of site frequency profiles than the empirical profile mixture models, we further simulated data for the LBA and LBR cases under the most complex available empirical mixture model: the LG+C60+F+Γ model (4 discrete gamma rates and alpha = 0.75) which had 61 frequency profiles. This allowed us to evaluate the relative performance of the various methods when the fitted mixture model (e.g., LG+C20+F+Γ) is much simpler than the generating model. A total of 345 sites were simulated for each of the C60+F components (the alignment has 21,045 sites) and 100 data sets were simulated for each pair of the long and short branches settings.

Simulation under 8-, 12-, 16-, and 20-taxon trees.—The foregoing simulated data sets were based on four taxon trees. As the number of taxa increases, more site-specific information is available for frequency estimation and performance of PMSF models can be expected to improve. To investigate, we simulated one LBR case where $a=0.4$ and $b=0.15$ under LG+C20+F+Γ for trees of 8-taxa, 12-taxa, 16-taxa and 20 taxa. The branch lengths *a* and *b* were so chosen as it was found to be difficult for the PMSF and MAX models to correctly estimate the right tree in 4-taxon simulation and LBR was rampant in this setting. An 8-taxon tree (Fig. 2) is obtained from the 4-taxon tree by bisecting four external edges at their midpoints. The A, B, C, and D taxa groups are similarly arranged in simulating under the 12-, 16-, and 20-taxa trees; as more taxa are added, they are added to a polytomy halfway along the external edges. In each simulation, ML scores were computed for the three topologies that differ by one split: AB|CD, AC|BD or AD|BC. Each topology was constrained to have taxa within groups of A–D appear together and, as in the generating model, internal edge lengths within the A–D groups were set to 0. All terminal edge lengths and other internal edge-lengths were estimated by ML. One hundred data sets were simulated for each generating setting and the ML scores for the three fixed topologies were compared to obtain counts of the number of times the correct trees had the largest score. Similarly, since the simulations showed (see the results below) that the LG+F+Γ model has substantial LBA bias under the LG+C20+F+Γ–simulated data for the 4-taxon trees, we investigated whether adding more taxa in the simulation would help alleviate the bias. We simulated 100 data sets each under 8-taxa, 12-taxa, 16-taxa and 20-taxa trees with the long branch *a* being 0.4 and the short branch *b* being 0.015. The generating trees are similar in structure to those in Figure 2 except that taxa group A has long branches and taxa group D has short branches so that the LBA bias can be induced. The LG+F+Γ, LG+C20+F+Γ, and PMSF models were applied to the simulated data to get the likelihood scores for the three constrained trees in a similar manner to the LBR case.

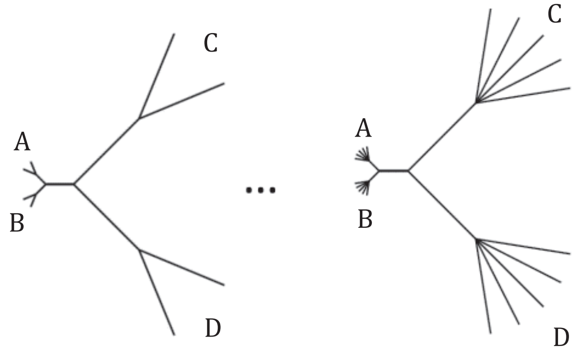


FIGURE 2. An 8-taxon tree (left) and a 20-taxon tree (right) used for simulating LBR. The branch lengths are not drawn to scale. The long branches have a length of 0.4 and the short branches have a length of 0.015. Simulating trees of 12 and 16 taxa have similar structure, adding more taxa to the groups of A, B, C, and D.

Simulation under structure-based mixture models.—The above simulated data are all based on a single amino acid replacement matrix—the LG rate matrix—with different site-frequency profiles. Le and Gascuel (2010) found that models accounting for solvent accessibility and secondary structures are highly beneficial in protein phylogenetics; these models were constructed by estimating both exchangeabilities and frequencies for sites falling into each of the various structural classes based on a large empirical data set of proteins with known structures. They derived six rate matrices: BUR-EXT: buried residues and extended structure; BUR-HEL: buried residues and helical structure; BUR-OTH, buried residues and other structures; EXP-EXT: exposed residues and extended structure; EXP-HEL: exposed residues and helical structure; and EXP-OTH, exposed residues and other structures. For the 4-taxon LBA and LBR cases (Fig. 1) we simulated 1000 sites under each of the six models and concatenated them into alignments of 6000 sites for each of the *a* and *b* settings. One hundred data sets were simulated for each setting. Tree searches were then conducted under the LG+F+Γ, EX_EHO+F+Γ, LG+C20+F+Γ, and PMSF models.

Empirical Data

Three hundred HSSP structure-sequence alignments.—In developing various site-heterogeneous models Le and Gascuel et al. used 300 sets of the HSSP protein alignments (Sander and Schneider 1994) as test data (Le et al. 2008b; Le and Gascuel 2010; Le et al. 2012). It is of interest to evaluate the performance of PMSF on these single protein data sets and to compare it with the previous results under different models. We applied the LG+ F+Γ, LG+C20+F+Γ, and PMSF models to the 300 data sets to estimate ML trees.

Five empirical phylogenomic data sets.—Five empirical multiprotein concatenated data sets are analyzed with the new models. The first "angiosperm" data set is a concatenation of 61 chloroplast protein sequences from *Amborella*, Nymphaea (water lilies) and 12 other land plants consisting of 24 taxa and 15,688 sites that was used to identify the deep splits in the angiosperm phylogeny (Leebens-Mack et al. 2005). In the latter study, a basal clade of *Amborella* + water lilies was inferred under ML for the nucleotide sequences, while ML estimation under JTT+I+Γ for the protein sequences placed *Amborella* alone as the deepest diverging taxon.

The second data set was from Brinkmann et al. (2005) and was made up of 24,291 aligned sites from 133 concatenated proteins. This data set had 40 taxa including a fast-evolving microsporidian species (*Encephalitozoon cuniculi*), 33 slow-evolving eukaryotic ingroup species and six archaea as outgroup taxa. When analyzed under single rate matrix models (e.g., JTT or WAG), instead of emerging in the correct position as a sister to fungi, *E. cuniculi* was positioned at the base of the eukaryotes, branching with the archaeal outgroup, apparently because the extremely long branch leading to *E. cuniculi* was attracted to the long branch separating the eukaryote ingroup and the archaeal outgroup. Brinkmann et al. (2005) showed that the LBA effect was reduced when the fastest-evolving Microsporidia proteins were gradually eliminated from the alignment. In the absence of the LBA bias, Microsporidia correctly branched with fungi.

The next two data sets are named based on the long-branched metazoan subgroups in the data sets that had controversial placements. One of these is the "nematode" data set (37 taxa 35,371 sites from 146 proteins) and the other is the "platyhelminth" (flatworms) data set (32 taxa 35,371 sites from the same 146 proteins). Lartillot et al. (2007) found the phylogenetic positions of these long-branched taxa depended on the outgroup when single-matrix models (e.g., WAG) were used. When fungal sequences were used as a distant outgroup of metazoans, the nematodes or platyhelminths branched with the outgroup to the exclusion of an arthropods plus deuterostomes clade (the so-called "Coelomata" group) as a result of an LBA artefact. However, when two choanoflagellates and a cnidarian were added to the outgroup to break up the long branch separating the ingroup and outgroup, nematodes and arthropods formed an "Ecdysozoa" clade (Aguinaldo et al. 1997) with deuterostomes splitting from other Metazoa at the deepest node. Similarly with the addition of the same three species to the fungal outgroup, platyhelminths and arthropods formed a "Protostomia" clade (Telford et al. 2015). Lartillot et al. (2007) showed that under a site-heterogeneous CAT+Poisson+Γ model (Lartillot and Philippe 2004), the Ecdysozoa or Protostomia clade was recovered for the two data sets respectively, regardless which of the two outgroups was used, effectively overcoming the LBA bias in both cases. It is interesting to see if the new PMSF models can also overcome the

LBA bias in the nematode and platyhelminth data sets and for this we will analyze the two more difficult cases where only the most distantly-related fungi are used as the outgroup.

The fifth data set we considered was that of Brown et al. (2013) which has 68 taxa and 43,615 sites from 159 proteins. Analyses of these data showed that two enigmatic microbial eukaryote lineages, the breviates and the apusomonads, grouped with animals and fungi (opisthokonts) to form a major eukaryotic assemblage called Obazoa (Brown et al. 2013). However Brown and colleagues showed that the precise position of the breviates within Obazoa depended on the phylogenetic model: ML and Bayesian analyses using single-matrix models (e.g., LG+F+Γ) recovered a clade of breviates (B) plus apusomonads (A) to the exclusion of opisthokonts (O) (the ((A,B),O) topology). Phylogenies inferred with the CAT+GTR+Γ model in Bayesian analysis (Brown et al. 2013), or ML analysis with site-class mixtures including LG+C20+F+Γ instead showed A and O as sister groups with the breviates at the base (the ((A,O),B) topology). The latter topology appeared to be most likely correct for the reasons discussed in Brown et al. (2013).

For each of the five data sets we first conducted model selection with IQ-TREE and found that whereas JTT amino acid replacement matrix fits best to the angiosperm data set (according to the Bayesian information criterion), the LG matrix is best for the other four data sets. We therefore conducted tree searches under JTT+F+Γ, JTT+C20+F+Γ, and JTT+PMSF+Γ for the angiosperm data set and LG+F+Γ and LG+C20+F+Γ and LG+PMSF+Γ for the other data sets. Two PMSF implementations were considered that differed in their guide trees. PMSF2 was based on the ML trees estimated under JTT (or LG)+F+Γ, while PMSF3 was based on the tree estimated under the C20+F mixture. Furthermore, for each of the five data sets we conducted the standard nonparametric bootstrap analysis (STNboot) each with 100 replicates and the ultrafast bootstrap analysis (UFboot; Minh et al. 2013) with 1000 replicates for the JTT (or LG)+F+Γ and JTT (or LG)+Γ+PMSF models. As nonparametric bootstrap analysis under C20+F or C60+F mixtures cannot be finished within a reasonable time, the UFboot approximation with 1000 replicates was performed for the mixture models. Finally, we also conducted Bayesian analyses under a CAT+GTR+Γ model with PhyloBayes-MPI (Lartillot et al. 2013) for the Angiosperm and Microsporidia data sets. Bayesian analyses under this model for the other data sets have already been carried out and published (Lartillot et al. 2007; Brown et al. 2013).

Results

Runtime and RAM Usage

To evaluate the relative running times of these various models in tree estimation, we used the Obazoa data set from Brown et al. (2013) to benchmark the time usage and memory required for ML tree estimation under several models in IQ-TREE (Table 2). The CPU time for tree searching under the LG+C20+F+Γ mixture model was 27.2 times that of the LG+F+Γ model. The LG+C60+F+Γ model spent 75.9 as much time as the latter. By contrast, PMSF2 required 3.3 times as much time as LG+F+Γ. It was 8.3 and 23.1 times faster than LG+C20+F+Γ and LG+C60+F+Γ, respectively. In terms of computer memory usage, the LG+C20+F+Γ mixture model used 20 times more RAM than the LG model while PMSF used only 22% more RAM than the LG model. As expected the LG+C60+F+Γ model used nearly three times of both RAM and computing time than the LG+C20+F+Γ mixture. The results indicate that PMSF has substantial computational advantages over the full mixture models.

Table 2. Running time and RAM usage for tree search on the Obazoa data (68 taxa 43,615 sites)

Models	CPU time	RAM (MB)
LG+F+Γ	5h:18m:55s	1768
LG+C20+F+Γ	6 days:0h:23m:3s	37,141
LG+C60+F+Γ	16 days:19h:12m:0s	107,888
LG+PMSF2+Γ	17h:29m:2s[a]	2160

Note: All the runs were conducted using a single core with IQ-TREE (version 1.5.1) on a computer cluster.

[a] PMSF required a small amount of time to initially fit guide tree under the LG+C20+F+Γ mixture and to get the PMSF profiles; this was counted in the CPU time shown. For this data set, this overhead is 14.5% of the total computing time for PMSF2.

Bootstrap Analysis Under PMSF

We tested three bootstrap analysis methods under PMSF, including: 1) the standard nonparametric bootstrap analysis that (a) refits LG+C20+F+Γ for each pseudoreplicate to get new bootstrap PMSF profiles for each of the resampled sites and (b) obtains the estimated bootstrap tree under PMSF; 2) a short-cut nonparametric bootstrap that does not refit LG+C20+F+Γ model for each pseudoreplicate but is otherwise the same as 1); and 3) an ultrafast bootstrap analysis (Minh et al. 2013). We applied the three methods to one HSSP data set with 19 taxa and 489 sites (Ord024_2h4m.all_gb.phy from Le and Gascuel 2010). The ML tree with the three sets of bootstrap values is shown in Supplementary Figure S1 available on Dryad at http://dx.doi.org/10.5061/dryad.gv1q5. The bootstrap values obtained from the two nonparametric bootstrap methods are nearly identical (P-values for all splits that had a different BP were >0.09 and therefore not significantly different, using a test of marginal homogeneity for paired samples; Kalbfleisch 1985). The ultrafast BPs, although correlated with the nonparametric BPs in general, were always larger and, for four splits, significantly differed from the nonparametric BPs (all P-values were <0.05 based on tests comparing two proportions

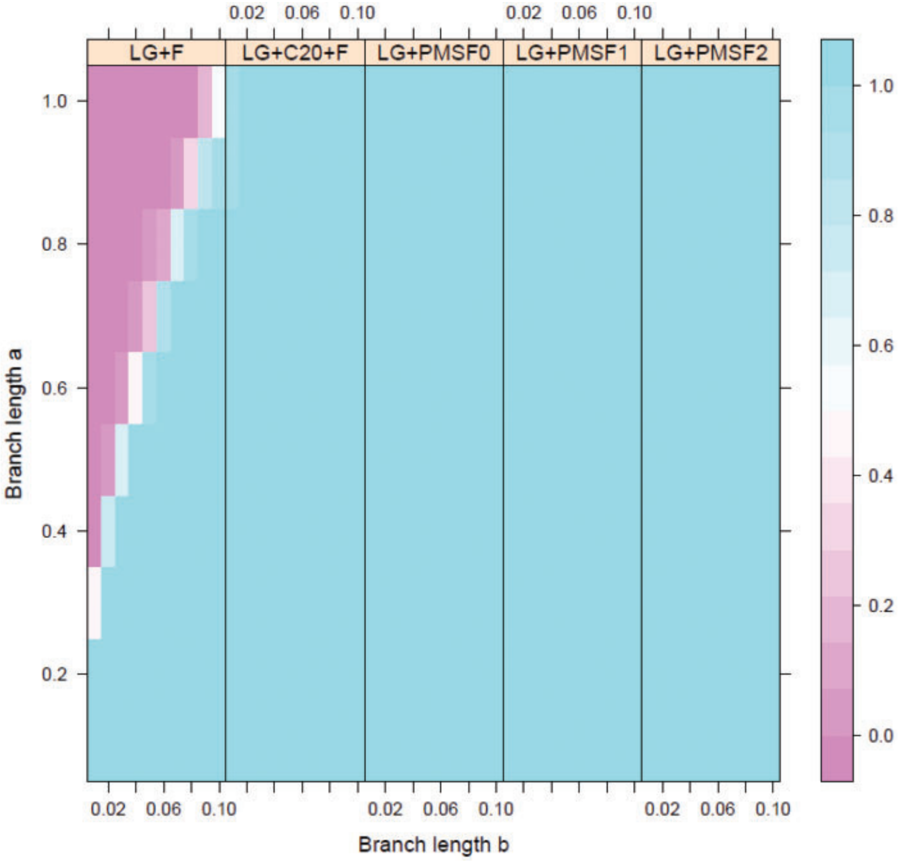


FIGURE 3. Proportions of correctly estimated trees under the various models in LBA simulations (21,000 sites in each data set) under LG+C20+F+Γ with "a" branches on the y-axis and "b" branches on the x-axis. All models contain +Γ which is omitted for title brevity.

of independent samples). Therefore, the short-cut nonparametric bootstrap approach is a valid and fast alternative to the standard nonparametric bootstrap approach and will be used throughout the paper. It is also implemented in IQ-TREE as the default nonparametric bootstrap method for PMSF.

Simulated Data

Under the LBA-simulation conditions.—For the data simulated under LG+C20+F+Γ for the LBA-inducing four taxon trees (Fig. 1 upper panel), ML trees were estimated under the various models and the proportions of the correctly estimated trees were plotted as a function of the branch lengths a and b on heat maps. Figure 3 shows the simulations for 21,000 sites in each data set. Under LG+F+Γ, the upper left area of the heat map shows an increasingly high proportion of incorrect trees as the long "a" branches increase and the short "b" branches decrease, revealing the typical LBA bias pattern. Tree estimations were 100% correct under the LG+C20+F+Γ mixture and PMSF, no matter which guide tree was used (Fig. 3). The MAX0 and MAX1 models also obtained 100% accuracy (figures not shown).

In Figure 3 the largest a/b ratio is 100 when $a=1$, $b=0.01$ and both the C20 mixture and PMSF models showed no LBA bias. When the a/b ratio gets even larger the proportion of correct estimations under the LG+C20+F+Γ model decreases but remains large under the PMSF (Table 3). In this setting, the LG+F+Γ model always estimated the LBA tree.

For the LBA simulation with much shorter alignments (1050 sites), the LG+C20+F+Γ, PMSF0, and PMSF1 models exhibit a relatively small LBA bias, and much less than that of LG+F+Γ (Supplementary Fig. S2 available on Dryad). The fact that the LBA is not present for PMSF for longer alignments suggests statistical consistency. In contrast, the LG model under these conditions, likely exhibits the LBA form of inconsistency.

LBA simulation under LG+F+Γ.—So far, the data were generated under the site-heterogeneous LG+C20+F+Γ model and, not surprisingly, the results showed that the correctly specified LG+C20+F+Γ mixture model outperformed LG+F+Γ in reducing LBA bias. On the other hand, the better performance of PMSF relative to LG+C20+F+Γ (Supplementary Figure S2 available on Dryad and Table 3), even when the incorrect guide tree was used was unexpected. It remains to be seen whether PMSF works as well for a simpler but standard simulation setting. Supplementary Figure S3 available on Dryad shows the results for data simulated

TABLE 3. The proportions of estimated correct trees and LBA trees for different a/b ratios in 4-taxon LBA data (21,000 sites) simulated under LG+ C20+F+Γ and estimated under LG+ C20+F+Γ, PMSF2 (the LG tree as the guide tree) and PMSF3 (the C20+F tree as the guide tree)

	a/b ratio	LG+ C20+F+Γ		PMSF2		PMSF3	
		Correct tree (%)	LBA tree (%)	Correct tree (%)	LBA tree (%)	Correct tree (%)	LBA tree (%)
$a=1, b=0.01$	100	100	0	100	0	100	0
$a=2, b=0.01$	200	58	42	97	1	98	1
$a=3, b=0.01$	300	39	59	90	1	89	1

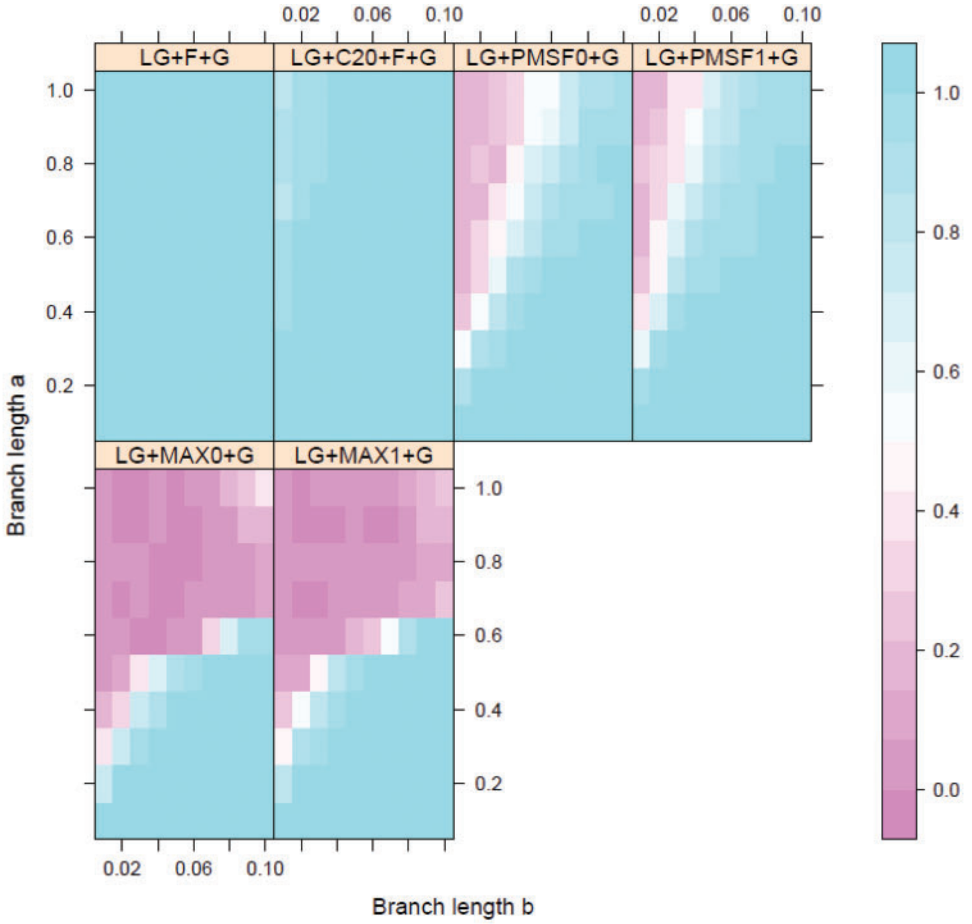


FIGURE 4. Proportions of correctly estimated trees under the various models in simulating LBR cases (21,000 sites) under LG+C20+F+Γ.

under LG+F+Γ when the ML trees were inferred under LG+F+Γ, LG+C20+F+Γ, and PMSF2. For the 1000-sites data all three models had a small LBA bias but the LG+F+Γ appeared slightly more biased than PMSF2 and C20+F, even though the LG model was the generating model. This suggests that PMSF and mixture models are more resistant to LBA than the LG model. When the sequence lengths increase (20,000 sites), all models always estimated the true tree (Supplementary Fig. S3 available on Dryad). As expected, as the sequences get longer, the estimated weights for the F component under the C20+F model get larger, effectively converging upon the simpler nested LG+F+Γ model; the estimated average F weight for the 1000-sites data was 0.85 ± 0.06 which increased to 0.95 ± 0.01 for the 20,000-sites data.

LBR-simulation conditions.—Figure 4 shows the heat maps of the proportions of correctly estimated trees for the various models under the LBR simulation conditions (Fig. 1 lower panel) for 21,000 sites. There is no LBR-bias in tree estimation under LG+F+Γ. By contrast the correct LG+C20+F+Γ model, which was used to generate the data, shows a mild LBR bias in the top left corner of the heat map (branch a is much longer than b). Both LG+PMSF0+Γ and LG+PMSF1+Γ models showed an even more pronounced LBR bias, and the MAX0 and

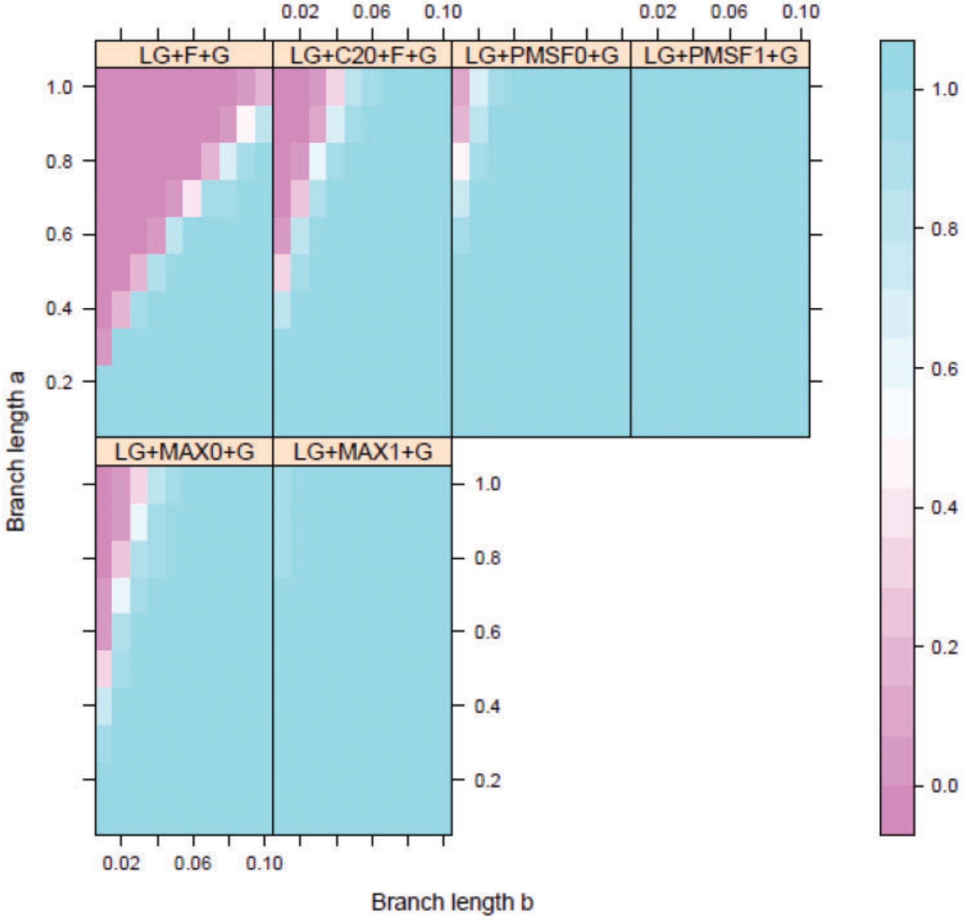


FIGURE 5. Proportions of correctly estimated trees under the various models in simulating LBA cases under LG+C60+F+Γ.

MAX1 models performed substantially worse. PMSF2 (i.e., using the LG+F+Γ tree as guide tree for PMSF estimation), was essentially equivalent to using the true tree as the guide tree, and had the same performance as PMSF1 in Figure 4 (not shown).

For the 4-taxon LBR simulation conditions, when the simulated sequences are shorter with 1050 sites, the LG+F+Γ model still shows no LBR bias but LG+C20+F+Γ and especially PMSF display this bias (Supplementary Fig. S4 available on Dryad) to a degree that is notably worse than for the 21,000-sites simulation (Fig. 4). That bias reduces as sequence length increases indicates that estimation with PMSF improves with the number of sites, which is suggestive (but not proof) of statistical consistency in this setting.

LBR simulation under LG+F+Γ.—We also simulated the data under LG+F+Γ for the LBR-inducing condition for 1000 sites and 20,000 sites respectively and estimated under LG+F+Γ, LG+C20+F+Γ and LG+PMSF2+Γ. For both sequence lengths (Supplementary Fig. S5 available on Dryad lower and upper panels), the LG model is the least LBR biased and the PMSF2 model is the most biased. However, for the longer sequences the LBR bias under all models is reduced.

Simulation with a more complex mixture (LG+C60+F+Γ) under LBA conditions.—In reality, almost every variable site in a protein family is subject to a unique set of structural and functional constraints that will manifest in a unique stationary amino acid frequency profile. To investigate the performance of smaller dimensional mixtures for cases in which the true simulating model is more complex as it is for real protein families, we simulated data under the 61 component LG+C60+F+Γ model and used simpler models for estimation over the branch length grid (Fig. 5). The LG+F+Γ model again shows a substantial LBA bias. The LG+C20+F+Γ is also LBA biased in this case, but to a lesser extent. Interestingly, PMSF1 shows no LBA bias and PMSF0 shows only a slight bias. PMSF2 shows the same results as PMSF0 (figure not shown). The MAX0 and MAX1 variants are slightly worse than the PMSF counterparts but still perform better than the LG+C20+F+Γ mixture model.

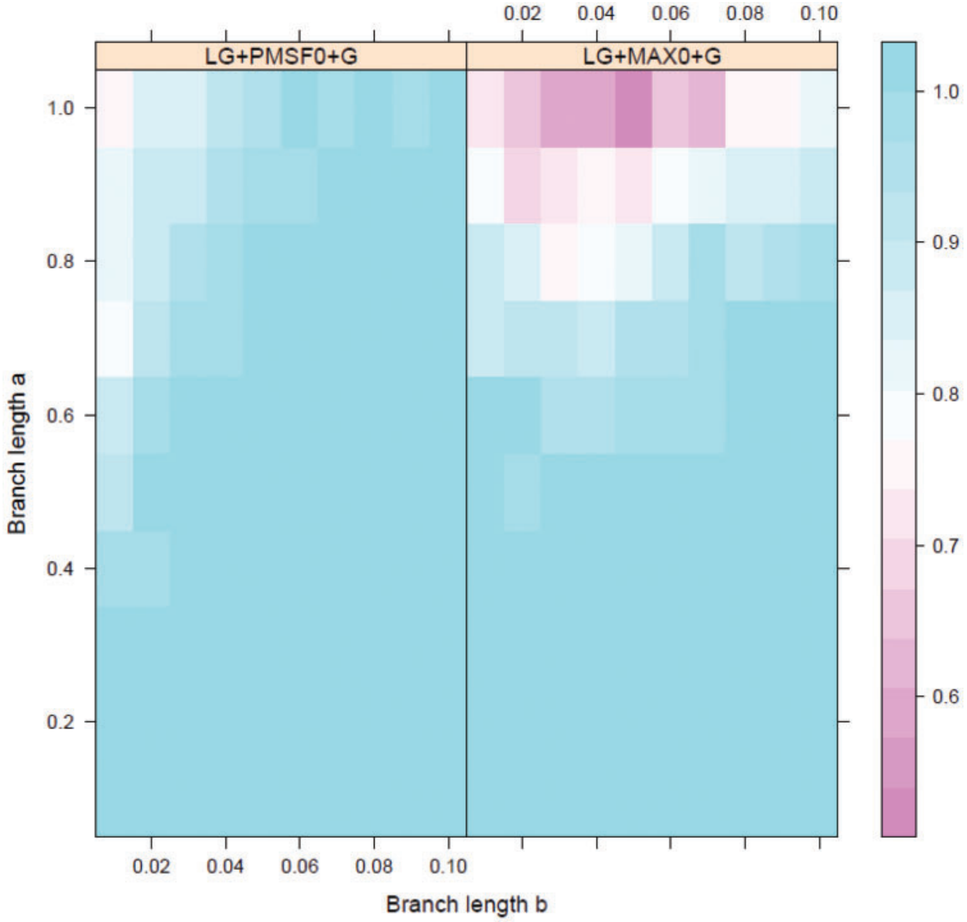


FIGURE 6. Proportions of correctly estimated trees in simulating LBR cases under LG+C60+F+Γ and estimated under PMSF0 and MAX0.

Simulation with a more complex mixture (LG+C60+F+Γ) under LBR conditions.—For the simulations employing LG+C60+F+Γ under LBR conditions, the estimated trees are 100% correct in all regions of the investigated parameter space when estimated under LG+F+Γ, LG+C20+F+Γ, PMSF1, and MAX1 (figures not shown). PMSF0 exhibits a small LBR bias which is more substantial under MAX0 (Fig. 6). The PMSF and MAX models are all derived under LG+C20+F+Γ with the correct guide tree (for PMSF1 and MAX1) or the wrong guide tree (for PMSF0 and MAX0). PMSF2 and MAX2 with the LG tree as guide tree, which is the true tree in all settings tested, also estimated 100% correct trees like PMSF1 and therefore showed no LBR bias in this simulation.

Taken together, for the data simulated under LG+C60+F+Γ, the PMSF1 model (based on the true guide tree and derived from the LG+C20+F+Γ mixture) performed best among all the models for both the LBA and LBR cases. Even PMSF0, based on the incorrect guide tree and derived also from LG+C20+F+Γ, showed less LBA bias than the LG+C20+F+Γ mixture model itself, although it did display some LBR bias. The performance of PMSF2 fell between PMSF0 and PMSF1: it showed a slight LBA bias as in PMSF0 and no LBR bias as in PMSF1.

Impact of taxonomic sampling on performance.—For the 4-taxon LBR generating tree with branch lengths $a=0.4$ and $b=0.015$ the proportions of the correctly estimated trees are all less than 8% under PMSF0, PMSF1, MAX0, and MAX1 (Fig. 4). Figure 7 shows that with increasing taxon sampling, the estimation accuracies increase rapidly. For the data (21,000 sites) simulated under the 8-taxon tree (see Fig. 2), PMSF1, PMSF0, MAX1, and MAX0 estimated the correct tree 82%, 38%, 62%, and 20% of the time. For the 12-taxon tree generated data, the percentages are 98%, 88%, 94%, and 56%. For the data generated under the 16-taxon and 20-taxon trees (Fig. 2) all estimated trees are 100% correct for all but PMSF0 which had an accuracy of 97% for the 16-taxon data sets. The result shows that increasing taxon sampling can effectively reduce the LBR bias for the PMSF and MAX models.

In Figure 3 and Supplementary Figure S2 available on Dryad the LG+F+Γ model is shown to have substantial LBA bias for the four taxon simulation. It is therefore of interest to know if increased taxon sampling will reduce

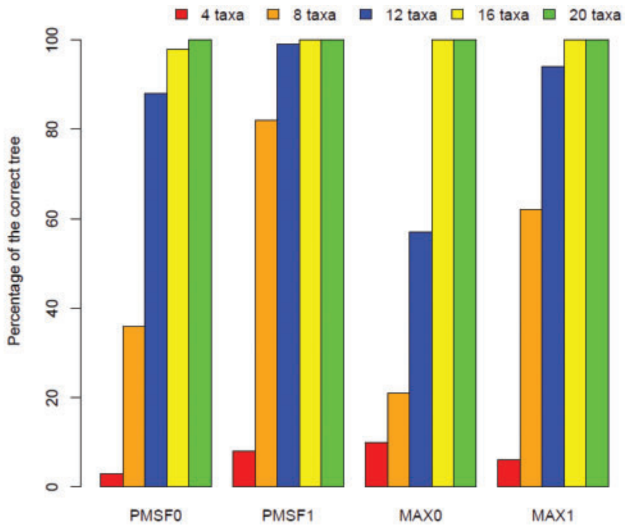


FIGURE 7. Simulation for the LBR setting (fixed branch lengths were $a=0.4$, $b=0.015$) for 4-taxon, 8-taxon, 12-taxon, 16-taxon, and 20-taxon trees with data generated under LG+C20+F+Γ and estimated under PMSF0, PMSF1, MAX0, and MAX1 (see Fig. 2).

this bias for the LG model. Table 4 shows that as the number of taxa increase the proportions of the estimated LBA trees reduce gradually. However the proportion of LBA trees is still at 14% for the data simulated under the 20-taxon tree.

All of the foregoing simulations show that PMSF0, PMSF1 and PMSF2 are relatively insensitive to the LBA bias in the ML estimation. They are even better than the C20+F mixture when the data sets are simulated under a larger number of amino acid frequency profiles (i.e., the 61 components under the C60+F mixture), even though the PMSFs, in this case, were still derived from the data fitted under the C20+F mixture. Compared with the LG+C20+F+Γ mixture model, PMSF and MAX models appear more prone to LBR bias. However, increasing the number of taxa in the data sets beyond 16–20 taxa, which is usually possible for empirical phylogenomic data, can effectively reduce the LBR bias. Since PMSF0 and PMSF1 performed better than MAX0 and MAX1 in all simulations we will consider only PMSF in the following studies.

Simulation under a mixture of the six structure-based amino acid replacement models.—Under LBA-inducing conditions, for the concatenated data simulated under BUR_EXT, BUR_HEL, BUR_OTH, EXP_EXT, EXP_HEL, and EXP_OTH, both the generating EX_EHO+F+Γ model and the LG+F+Γ model displayed a small LBA bias when the ratio of the long branch lengths to short branch lengths is large. The LG+C20+F+Γ model showed much less bias than the LG model and the PMSF2 model was the least biased among the four models (Supplementary Fig. S6 available on Dryad lower panel). Indeed, except in one case where $a=1.0$ and $b=0.01$, an LBA tree was estimated in one out of 100 simulations, PMSF achieved 100% correct estimations in the other 9999 simulations under the 100 pairs of the a and b settings. Therefore, these simulations also demonstrate that the PMSF and C20+F mixture models can effectively alleviate the LBA bias for the data generated under several structure-based rate matrices.

Under the LBR-inducing conditions, the LG+F+Γ model had no bias, EX_EHO+F+Γ had a very slight bias, and LG+C20+F+Γ showed a greater bias. LG+PMSF2+Γ had a substantial LBR bias (Supplementary Fig. S6 available on Dryad upper panel). The simulated data have only four taxa and 6000 sites. Adding more taxa or increasing sequence length is expected to reduce the LBR bias. For instance, the top left corner in Figure S6 corresponding to branch lengths $a=1.0$ and $b=0.01$ had the lowest estimation accuracy (25%) for LG+PMSF2+Γ in the 4-taxon LBR simulation. When simulating under EX_EHO under an 8-taxon tree (see Fig. 2) the proportion of correctly estimated trees under PMSF increases to 54% and it further reaches 83% when a 16-taxon tree was used for simulation. This suggests that PMSF performance under these conditions substantially improves with better taxonomic sampling. However, as discussed in a later section ("Comparison with other models"), simulations under the plain LG or EX_EHO models have questionable relevance to the performance of PMSF with real data; these models typically fit real data worse than sufficiently complex site-specific profile mixture models such as C20+F.

Empirical Data

The 300 HSSP test data sets.—The LG+F+Γ, LG+C20+F+Γ, LG+PMSF2+Γ and LG+PMSF3+Γ models were used to infer the ML trees from the 300 data sets. Using the Robinson–Foulds distance (Robinson and Foulds 1981) to measure the topology difference between the four models, Table 5 (and see Supplementary Materials S2 available on Dryad for more details) indicate the site-heterogeneous models (C20+F mixture and PMSF) estimated different topologies from the site-homogenous LG model in the majority of the cases. Similar findings were reached when comparing LG with the other site heterogeneous models (including LG4M, LG4X and the structure models) (Le and Gascuel 2010; Le et al. 2012). All three site-heterogeneous models estimated trees more similar to one-another than to the tree estimated under the LG+F+Γ model. The C20+F mixture model tends to estimate trees more similar to PMSF3 than to PMSF2, indicating that the guide tree has an impact on the estimation.

Comparing the log-likelihoods of the estimated trees under the models, 277 data sets show significant evidence for C20+F over LG based on a likelihood ratio (LR) test, while the other 23 data sets do not yield significant results. It should be noted that the LR test is expected to be conservative under these conditions (i.e., the true P-values are smaller than those yielded by χ^2 with 20 degrees of freedom) because for the C20+F model to be reduced to LG, 20 of the mixture weight

TABLE 4. The proportions of estimated correct trees and LBA trees for simulated data with increased taxon sampling

	LG+F+Γ		LG+ C20+F+Γ		PMSF2	
	Correct tree (%)	LBA tree (%)	Correct tree (%)	LBA tree (%)	Correct tree (%)	LBA tree (%)
4 taxa	0	100	100	0	100	0
8 taxa	18	82	100	0	100	0
12 taxa	48	52	100	0	100	0
16 taxa	83	17	100	0	100	0
20 taxa	86	14	100	0	100	0

Note: One hundred alignments of 21,000 sites were generated respectively under LG+C20+F+Γ for LBA-inducing trees of 4, 8, 12, 16, and 20 taxa with the long branch $a=0.4$ and short branch $b=0.015$.

TABLE 5. Topology estimation for the 300 HSSP test data sets

Models in comparison		Number of times same topology was estimated	Average RF distance over 300 data sets
LG+F+Γ	LG+C20+F+Γ	38	15.13
LG+F+Γ	LG+PMSF2+Γ	51	11.36
LG+F+Γ	LG+PMSF3+Γ	38	15.31
LG+C20+F+Γ	LG+PMSF2+Γ	103	8.07
LG+C20+F+Γ	LG+PMSF3+Γ	198	2.31
LG+PMSF2+Γ	LG+PMSF3+Γ	114	7.87

parameters must be set to 0, which is on the boundary of their parameter space (Self and Liang 1987). As a result, some of the 23 data sets that were not rejected using the χ^2 critical value, are likely truly significantly better fit by the C20+F model. Interestingly, the mean number of taxa and number of sites are 45.02 ± 22.87 and 244.09 ± 141.28 respectively in the 277 significant data sets, while the corresponding statistics are 16.61 ± 5.98 and 282.17 ± 206.37 respectively for the 23 nonsignificant ones. A t-test shows that the average number of taxa is significantly higher ($P<10^{-15}$) in the 277 data sets than the other 23 data sets, while the average number of sites is not significantly different ($P=0.39$). We suggest that information about frequency heterogeneity across sites increases as the number of taxa increases—more so than as the number of sites increases. This likely explains why the 277 data sets tended to show stronger evidence of a better fit of C20+F versus LG than the remaining 23 data sets.

Another measure of the relative performance of the models is the Akaike information Criterion (AIC) (Akaike 1974). Le and Gascuel (2010) found that the EX_EHO model had per site AIC gains (i.e., smaller AIC) over the LG model for almost all 300 HSSP data sets (see Fig. 2 in Le and Gascuel, 2010). It is interesting to add the LG+C20+F+Γ mixture model in this comparison. The likelihood under the C20+F mixture is always greater than EX_EHO, which is in turn almost always greater than LG. In terms of the AIC, the C20+F mixture is lower than LG and EX_EHO in 270 and 273 data sets, respectively. We calculated, for each model, the average per site AIC for the 300 data sets combined, using Equation (6) in Le and Gascuel (2010). The per site AIC improvement of EX_EHO versus LG is relatively small at 0.20 while that improvement of C20+F versus LG is much larger at 1.15 per site. This indicates the C20+F mixture fit the data much better than both LG and EX_EHO.

Likelihood ratio tests and AIC cannot directly be used to compare the C20+F and PMSF models, as the "parameters" of the latter model are not estimated by ML. However, in comparing the log-likelihoods between PMSF2 and PMSF3, we found that PMSF3 estimated higher likelihoods in 219 data sets, PMSF2 did better in 70 data sets and both had the same likelihoods for 11 alignments. This suggests that using the C20+F guide tree (as in PMSF3) rather than the LG guide tree (PMSF2) may improve estimation under PMSF.

Altogether, the analyses of the 300 HSSP data suggest that topologies estimated within the class of site-heterogeneous models are more likely to be similar to each other than to the topology estimated under the LG model. Although it is largely unknown what the true topologies are for these data sets, as pointed out in Le and Gascuel (2010), these topology and likelihood differences should be of interest to the phylogeneticists studying these proteins.

Deep angiosperm phylogeny.—For the angiosperm concatenated chloroplast-encoded protein data set (Leebens-Mack et al. 2005), we conducted tree topology searches under the three models: JTT+F+Γ, JTT+C20+F+Γ mixture model, and JTT+PMSF2+Γ (i.e., JTT+F+Γ tree as the guide tree and derived PMSF profile under JTT+C20+F+Γ). As in Leebens-Mack et al. (2005), the ML tree under JTT+F+Γ put *Amborella* as the sister of a clade made up of all other flowering plants, although the bootstrap support is only 49%

TABLE 6. Bootstrap support values (%) for the correct Microsporidia+Fungi (M+F) split and incorrect Microsporidia+Archaea (M+A) split for the Microsporidia data set

	M+F split		M+A split		
	UFboot[a]	STNboot[a]	UFboot	STNboot	CPU time for standard bootstrap
LG+F+Γ	2	2	98	98	610 h 46 min
LG+C20+F+Γ	63	60[b]	37	40[b]	17,175 h 38 min[b]
LG+PMSF2+Γ	77	82	23	18	2194 h 50 min
LG+PMSF3+Γ	100	100	0	0	2223 h 55 min

[a] UFboot is ultrafast bootstrap support; STNboot is standard nonparametric bootstrap support.

[b] For running standard nonparametric bootstrapping under LG+C20+F+Γ, due to computer resource constraints, only 43 data replicates were completed which took 7385.5 CPU h, which extrapolates to 17,175 h 38 min for standard 100 bootstrap replicates. The STNboot values are based on the 43 bootstrap replicates.

(Supplementary Fig. S7 available on Dryad). In contrast, the two ML trees estimated under the JTT+C20+F+Γ and PMSF2 models recovered an *Amborella* plus water lilies (*Nuphar* and *Nymphaea*) as a clade (The C20+F UFboot = 91%, PMSF2 STNboot = 90%) sister to the remaining angiosperms (the latter clade had the C20+F UFboot = 100% and the PMSF2 STNboot = 100%) (Supplementary Figs. S8 and S9 available on Dryad). Finally, analyzing the data set under CAT+GTR+Γ also grouped *Amborella* and the water lilies in a clade with 0.94 posterior probability (PP) with the "remaining angiosperm" group gaining PP = 0.99 (Supplementary Fig. S10 available on Dryad). Although the position of *Amborella* appears to be still under active debate (Drew et al. 2014; Wickett et al. 2014; Goremykin et al. 2015), the site-heterogeneous models converge on *Amborella* + water lilies forming a basal angiosperm clade based on this data set.

The placement of Microsporidia in the tree of eukaryotes.—Tree topology searches were conducted under the five models: LG+F+Γ, LG+C20+F+Γ, LG+PMSF2+Γ, LG+PMSF3+Γ, and CAT+GTR+Γ. As expected, the LG+F+Γ model recovered an incorrect tree with the LBA bias where Microsporidia is grouped with archaea to the exclusion of other eukaryotes (STNboot = 98%, Supplementary Fig. S11 available on Dryad). The LG+C20+F+Γ mixture model estimated a topology where Microsporidia and Fungi form a clade (Supplementary Fig. S12 available on Dryad). Tree searching under the LG+PMSF3+Γ model recovered the same topology (Supplementary Fig. S13 available on Dryad). The bootstrap support under PMSF3 for the *E. cuniculi* + fungi split was 100% using both STNboot and UFboot. The optimal tree estimated with LG+PMSF2+Γ (Supplementary Fig. S14 available on Dryad) slightly differs from the tree estimated under LG+PMSF3+Γ but the *E. cuniculi* + Fungi group was also recovered, receiving 77% UFboot and 82% STNboot support respectively. This demonstrates that even if an incorrect LBA-biased topology is used as the guide tree to derive PMSF, it is possible to estimate the correct topology avoiding the LBA bias under this model. The CAT+GTR+Γ model also recovered a tree with an *E. cuniculi* + Fungi split (PP = 1.0) (Supplementary Fig. S15 available on Dryad).

Table 6 shows the bootstrap supports for the correct Microsporidia + Fungi (M+F) split and the incorrect Microsporidia + Archaea (M+A) split for the various models. The data indicate that both PMSF2 and PMSF3 have higher support for the correct M+F split than the LG+C20+F+Γ mixture model, while the LG+F+Γ model give very high support for the incorrect M+A split. Table 6 also gives the CPU time used for the 100 standard bootstraps under the LG+F+Γ, LG+C20+F+Γ, and PMSF models. Runs using the two PMSF models spent 3.3 and 3.6 times as much time respectively as the LG model, whereas they were 7.7 and 8.6 times faster than the C20+F mixture model. Since the STNboot analysis is extremely computationally demanding under the C20+F mixture, the UFboot approximation was used for the remaining empirical data analyses.

The placement of nematodes and platyhelminths in the metazoan tree.—For the nematode data, tree searching under LG+F+Γ yielded the same topology (Supplementary Fig. S16 available on Dryad) as the WAG+Γ model in Lartillot et al. (2007), namely arthropods formed a clade with deuterostomes (i.e., the so-called "Coelomata" topology) with nematodes splitting off earlier (i.e., joining the long branch leading to the outgroup). Analyses based on the LG+C20+F+Γ mixture however successfully avoided the LBA bias and supported the widely accepted Ecdysozoa clade (nematodes + arthropods) (Supplementary Fig. S17 available on Dryad). We derived the PMSF2 and PMSF3 models under the LG+C20+F+Γ mixture based on the incorrect LG tree and correct LG+C20+F+Γ tree respectively. Tree searching using PMSF2 and PMSF3 both recovered the same correct topology with the Ecdysozoa clade (Supplementary Figs. S18 and S19 available on Dryad). The STNboot support for this clade is 65% in PMSF2 and 100% in PMSF3.

For the ML trees estimated under the various models the split between the fungal outgroups and the in-group taxa (nematodes, arthropods and deuterostomes) had a bootstrap support of 100% for both STNboot and

TABLE 7. Bootstrap support values (%) for the position of the nematodes within Metazoa

	nematodes + arthropods (Ecdysozoa)		deuterostomes + arthropods (Coelomata)	
	UFboot	STNboot	UFboot	STNboot
LG+F+Γ	21	18	79	82
LG+C20+F+Γ	76	NA	24	NA
LG+PMSF2+Γ	69	65	31	35
LG+PMSF3+Γ	100	100	0	0

NA = data not available (too computational demanding under the model).

UFboot, but the support values within the three in-groups were different between the models. Table 7 lists the bootstrap support values for the correct Ecdysozoa split (nematodes + arthropods) and incorrect Coelomata split (deuterostomes + arthropods) under the various models. PMSF3 shows the highest support for the Ecdysozoa group, while LG+C20+F+Γ and PMSF2 have relatively lower support for this clade. The LG model gives relatively high support for the incorrect Coelomata clade.

For the platyhelminth data set, we conducted similar analyses to get ML trees under LG+F+Γ, LG+C20+F+Γ, PMSF2, and PMSF3. All analyses yielded the same incorrect topology displaying the Coelomata clade, with the long-branched platyhelminths clustering with the outgroup (Supplementary Fig. S20 available on Dryad for the ML tree under LG). Under the assumption that the LG+C20+F+Γ model may not have a rich enough set of classes to accurately model this data set, we fit the 61 component LG+C60+F+Γ model and searched for the ML tree. However, this yielded the same incorrect Coelomata topology. Interestingly, tree searching using PMSF derived from the foregoing LG+C60+F+Γ mixture instead recovered a tree with the correct Protostomia clade, albeit with weak STNboot support at 55% (Supplementary Fig. S21 available on Dryad).

Similar to the nematode data, the split between the fungal outgroups and the in-group taxa (platyhelminth, arthropods and deuterostomes) had 100% support for both STNboot and UFboot under the various models, but the supports within the three in-groups were different among the models. Table 8 shows the bootstrap support values for the Protostomia and Coelomata splits for the six models used in the ML tree searches. This again shows that only PMSF based on the LG+C60+F+Γ mixture gives a higher support for the Protostomia split than the Coelomata split and it is true even though the guide tree used to obtain the PMSF profile has a Coelomata split.

The placement of breviates and apusomonads in the tree of eukaryotes.—We conducted ML tree searches for the data under the LG+F+Γ, LG+C20+F+Γ, PMSF2, and PMSF3 models. As in Brown et al. (2013) the LG+F+Γ model estimated an Obazoa clade with the opisthokonts at its base and breviates and apusomonads are sister to each other forming a ((A,B),O) topology. All the other models (Supplementary Figs. S22–S24 available on Dryad) recovered an Obazoa clade with the correct ((A,O),B) topology. The STNboot support for the (A,O) split is 78% and 92% for PMSF2 and PMSF3 respectively and the support for the Obazoa clade ((A,O),B) is 99% and 100% respectively for the two models.

TABLE 8. Bootstrap support values (%) for the placement of platyhelminths within Metazoa

	platyhelminth + arthropods (Protostomia)		deuterostomes + arthropods (Coelomata)	
	UFboot	STNboot	UFboot	STNboot
LG+F+Γ	0	0	100	100
LG+C20+F+Γ	31	NA	69	NA
LG+PMSF2+Γ	45	55	55	45
LG+PMSF3+Γ	47	46	53	54
LG+C60+F+Γ	38	NA	62	NA
LG+PMSF.C60[a]+Γ	62	55	38	45

NA = data not available (too computational demanding under the model).
[a]PMSF.C60: the frequency profile is derived under LG+C60+F+Γ based on the LG+F+Γ tree.

TABLE 9. Bootstrap support values (%) for the placement of the breviates within Obazoa

	apusomonads + opisthokonts		apusomonads + breviates	
	UFboot	STNboot	UFboot	STNboot
LG+F+Γ	0	2	100	98
LG+C20+F+Γ	95	NA	5	NA
LG+PMSF2+Γ	85	78	15	22
LG+PMSF3+Γ	97	92	3	8

Table 9 shows the bootstrap supports for the two types of configuration of the Obazoa clade for the above models used in the ML tree searches. The results again indicate PMSFs gives relatively higher support for the correct ((A,O),B) topology than the incorrect ((A,B),O). In particular for PMSF2 which were derived from the ML tree with a ((A,B),O) split, the (A,O),B) topology was estimated under the model.

DISCUSSION

Comparisons of PMSF with Simpler Mixture Models

Our new PMSF method performs as well, if not better, than C20+F and C60+F empirical mixture models in estimating phylogenies in the presence of site-heterogeneity both in simulations and in empirical settings, yet they are much more computationally efficient. However, several other empirical mixture models including CF4 (Wang et al. 2008), LG4X

(Le et al. 2012), and the structure-based EX_EHO (Le and Gascuel 2010) contain fewer mixture components and therefore are also quite computationally efficient compared with the C20+F and C60+F mixtures. It of interest, therefore, to know whether they perform as well as the more complex mixtures and PMSF in tree estimation in empirical settings. We checked whether these simpler mixture models are able to overcome the LBA biases in the empirical phylogenomic data sets. In the Supplementary Materials available on Dryad (page 27), we show the relevant splits and their UFboot support values for the ML trees estimated under LG+CF4+F+Γ, LG4X+Γ, and EX_EHO+F+Γ for the five empirical data sets. For the angiosperm data, these three models estimated an *Amborella* + water lilies clade with 64%, 62%, and 64% UFboot support respectively. The same split was also estimated under the C20+F, PMSF and CAT-GTR models (Supplementary Figs. S8–S10 available on Dryad) but with higher support values (all >91%). For the remaining four empirical data sets, the CF4+F and EX_EHO models estimated topologies displaying the "long branch attraction" groupings described above (with UFboot support >66% in all cases). The LG4X model also estimated the "LBA" split for the Microsporidia, the platyhelminth and the Obazoa data sets with high support (UFboot >92% in all cases). On the other hand, for the nematode data set, the LG4X estimated the correct Ecdysozoa clade with modest support (UFboot = 73). Overall, however, the simpler mixture models like EX_EHO, LG4X, and CF4+F are less effective at overcoming the LBA bias in real phylogenomic data sets.

Fit of Models to Real data and Realism in Simulations

In the above analysis we showed that the average per site AIC gain relative to the LG model for the combined 300 HSSP protein data sets are much larger under the C20+F mixture than EX_EHO. This is also true for the five phylogenomic data sets considered here (Supplementary Table S1 available on Dryad). These results indicate that the C20+F mixture model fits empirical data—either single protein or multi-protein alignments—better than the LG or EX_EHO models in almost all cases. Thus the C20+F mixture model better captures the site-specific nature of the substitution process than the other two models. This suggests that the performance of the various models considered here in simulations under site-specific amino acid frequency distributions like the C20+F and C60+F mixtures (Figs. 3–7) are more relevant to the performance of these models on real data than the simulations under the LG or the EX_EHO models (e.g., Supplementary Figs. S3, S5, and S6 available on Dryad).

Regardless, the fact that PMSF outperform LG, EX_EHO and even C20+F (for the C60+F-simulated data) for the LBA simulation cases indicate that this new model can be effective overcoming the long branch attraction artefact, one of the most difficult biases in phylogenetic inference (Philippe *et al.* 2011). PMSF does display a LBR bias in the 4-taxon LBR-simulation cases. However our results indicate that the LBR bias can be remedied through increasing taxon sampling and longer sequences.

The Issue of Model Selection under PMSF

Direct comparison of likelihoods from PMSF with other models is complicated by the nature of fitting under PMSF. Traditional approaches to model selection like AIC, BIC, or LR tests require counts of the number of parameters estimated in the models under consideration and assume ML estimation was used in fitting. The number of estimated parameters for PMSF is not clearly defined. A naïve approach is to count 19 parameters at each site for the estimated amino acid frequencies in the PMSF at that site. This would be appropriate if frequencies were estimated by ML separately at each site. However, PMSF, which uses empirical Bayes estimates of frequencies, is much more restrictive in its estimation. To quantify this, consider the relative volume of the space of all frequencies at a site, to the volume of the space of frequencies constructed from mixtures of C20+F frequencies, which contains the space of allowable PMSF frequency vectors at any site. The qhull software of Barber et al. (1996) can be used to calculate such volumes and gives that the volume of the full space of frequencies is roughly 2×10^{20} times as large as that of the space of mixtures of C20+F frequencies.

Another illustration of the pitfalls of counting frequencies as parameters in PMSF can be seen by considering the use of LR tests for model selection. For standard model comparisons involving nested models, under the null hypothesis that simpler model is correct, the LR statistic (twice the log-likelihood difference between a more complex model and a simpler model) is expected to be χ^2 distributed with degrees of freedom equal to the differences in the number of parameters between the models. So in this context, if PMSF were approximately the same as ML estimation of 19 frequencies per site, then the LR statistic for a PMSF/single-frequency model (e.g., JTT+F+Γ) test, if the latter simpler model were correct, should be χ^2 distributed with 19× (No. of sites − 1). To test this, we generated a parametric bootstrap distribution of the LR statistics between a PMSF model (fit using JTT+C20+F+Γ) and the JTT+F+Γ model applied to 100 sets of the simulated angiosperm data under the JTT+F+Γ model and the ML tree (see the Methods in Supplementary Fig. S25 available on Dryad). Supplementary Figure S25 available on Dryad shows the bootstrap distribution. The critical value for a LR test with an α-level of 0.01 from this distribution is 2342 and the observed LR statistic for this data set is 26,150. Since the observed LR statistic is much larger than both the mean and 99th percentile of the bootstrap distribution (being 1400 and 2342 respectively), the first conclusion that can be drawn is that PMSF has significantly improved model fit relative to JTT+F+Γ. The second conclusion is that standard model comparison based on differences in the number of parameters are inappropriate. Whereas the appropriate critical value coming from the bootstrap distribution is 2342, the 99th percentile of a χ^2 distribution with 12,548 $\times$ 19 = 238,412 degrees of freedom is 240,021, which is more than 100 times greater than the true critical value. Thus the increased "flexibility" of PMSF relative to the single frequency model is far lower than 19 additional free parameters per site.

It is clear that the simple application of an LR test or other model selection methods assuming 19 free parameters per site cannot be validly used to compare PMSF to single frequency models. Although the parametric bootstrapping approach described here is a reasonable alternative, it is computationally intensive and development of new model selection criteria appropriate for this setting will be an important avenue for future investigations.

Comparing PMSF Models Under Different Guide Trees

It is usually the case that likelihoods for models of the same dimension are comparable and that the model with the largest likelihood is expected, with large numbers of observations, to be the correct one. Comparisons can become problematic when the numbers of parameters fit under one model are greater than under another model or, stated more generally, if there is greater flexibility in estimating parameters under one model than another. Since PMSF profiles constructed from two different guide trees have the same flexibility of parameter estimation, the expectation is that their likelihoods will be comparable. Thus the tree giving the largest log-likelihood is the one to be preferred. The reason for preferring the tree with the largest log-likelihood stems from the fundamental reason that tree estimation is statistically consistent with a fixed frequency vector: the expected log-likelihood for the true tree and frequencies, will always be larger than the expected log-likelihood for any other tree and frequencies. For frequencies generated with a finite number of classes, this implies that for any fixed class, the sum of site log-likelihoods for the correct tree and correct frequency profile for that class will eventually be larger than the sum of site log-likelihoods for any particular alternative tree and profile. Since this result applies for each class and since the overall log-likelihood is the sum of the site log-likelihoods over all classes, it follows that the log likelihood for the true tree and correct frequency profiles will eventually be larger than for any other tree. Since correct frequency profiles are not exactly known, PMSF estimates them and, for the correct tree, provides an approximation to the log-likelihood for the true tree and correct frequency profiles, which is the reason for preferring trees giving largest log likelihoods. Therefore, if different PMSF models are derived from the same mixture model with different guide trees or from different mixture models of the same dimension, the log-likelihoods of a given topology and the accompanying PMSF model can be directly compared. This can be used to select the best topology in the case that multiple alternative optimal topologies have been recovered based on tree-searching using different PMSF models. On the other hand, log-likelihoods of trees estimated under PMSF models derived from mixture models of different dimensions (e.g., PMSFs from LG+C20+F+Γ vs. LG+C60+F+Γ) cannot be straightforwardly compared, nor can standard model selection frameworks be applied. This is because, as empirical Bayes posterior mean estimates, the PMSFs are fitted to the individual sites. Therefore, the more complex the mixture models upon which they are based, the more site classes and class weights are being used to derive these posterior means and the greater the range of frequencies that can be fit. So the more complex mixtures can be expected to give frequencies that are closer to the maximum likelihood frequencies at the site.

Developing Better PMSF Profiles

Although for the five empirical phylogenomic data sets the PMSF profiles derived from the incorrect guide trees (i.e., the LG or JTT trees in the examples) are still able to estimate the correct topologies, it is ideal to use more accurate trees, whenever available, as guide trees to obtain the PMSF profiles. This will increase the chance of estimating accurate topologies, as the LBA and LBR simulations show, and lend higher bootstrap supports for the correct splits (Tables 6–9). For new phylogenomic data without well-accepted phylogenies, one may first estimate a ML tree under LG+F+Γ, which is computationally fast, and use this tree as a guide tree to fit a LG+C20+F+Γ or (even better LG+C60+F+Γ model) to obtain the PMSF profile. Then tree searching can be completed under LG+PMSF+Γ to get an ML tree, which is relatively faster than estimating a tree under the full LG+C20+F+Γ mixture. Alternatively one might iteratively update the guide tree. As above, this approach would use PMSF derived from the LG+F+Γ guide tree to obtain a first LG+PMSF+Γ tree. That LG+PMSF+Γ tree would then be used as a new guide tree, leading to new PMSF and hence a new LG+PMSF+Γ tree. The process could be repeated a fixed number of times or until no further topological changes occur. As the guide tree can be expected to get more accurate with each step, the PMSF profile can be expected to get closer to the true profile, which should allow more accurate phylogenies to be estimated.

In the PMSF approach we investigated the use of posterior mean and posterior maximum frequencies, settling upon the former as the better choice. While it seems clear that using observed site frequency profiles in place of the posterior mean frequencies will give rise to undesirable properties due to sparseness concerns, it is possible that approaches that smooth frequency estimates in some way, through for instance the addition of pseudocounts may give reasonable alternative frequency profiles for the general approach here. Such modifications likely would still require relatively large numbers of taxa whereas posterior means smooth the frequencies to a greater degree.

In summary, we proposed and implemented PMSF models to approximate the empirical profile mixture models. A PMSF is a conditional mean frequency vector estimated for each site based on a fitted mixture model and a guide tree. These models provide substantial computational savings compared with mixture models. The computational efficiency of PMSF makes them tractable to analyze large phylogenomic data sets that are continually growing in numbers of genes/proteins and taxa sampled. They also allow for standard nonparametric bootstrap analysis, as shown in all five empirical case studies. Our simulations and empirical data analyses demonstrate that PMSF can effectively ameliorate LBA artefacts and, in a few cases, they provide more accurate topological estimates than the mixture models themselves.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.gv1q5.

Funding

This work was supported by Discovery grants awarded to A.J.R. and E.S. by the Natural Sciences and Engineering Research Council of Canada. A.J.R. acknowledges support from the Canadian Institute for Advanced Research Program in Integrated Microbial Diversity and the Canada Research Chairs program. B.Q.M. was partly supported by the Austrian Science Fund (FWF I-2805-B29).

Acknowledgments

The authors thank Oliver Gascuel, Mark Holder, Nicolas Lartillot, Paul Lewis, Tal Pupko, and an anonymous reviewer for constructive comments on the paper.

Appendix: The Relative Computational Cost of Likelihood Evaluation

For PMSF, and a fixed rate and site, the likelihood contribution, L is obtained through a sequence of updates. For some fixed rooting of the tree, let $L_i(p)$, denote the conditional probability of the data at descendent terminal nodes, given amino acid i at parent node p. The partial likelihood $L_i(p)$ is obtained a product of partial likelihoods, $L_i(p,e_1)L_i(p,e_2)$ over two descendent edges, where $L_i(p,e)$ gives the conditional probability of descendent data from node p and along edge e. For a sequence of internal nodes in an arbitrarily rooted tree, partial likelihoods, or more precisely, conditional probabilities, $L_i(p,e)$ of descendent data, given amino acid i at parent nodes, p, are successively updated along their descendent branches through

$$L_i(p,e) = \sum_{k=1}^{c} U_{ik}^{-1}\exp(\lambda_k t)U_{ks(d)}, \qquad \text{external edge} \tag{A.1}$$

$$L_i(p,e) = \sum_{j=1}^{c}\sum_{k=1}^{c} U_{ik}^{-1}\exp(\lambda_k t)U_{kj}L_j(d), \quad \text{internal edge} \tag{A.2}$$

where c is the number of character states (20 for amino acids). Here U, U^{-1} and λ_k come from the eigen-decompositions of the rate matrices and t denotes the edge-length. In the external edge calculation, $s(d)$ denotes the amino acid at the terminal node and in the internal edge calculation $L_j(d)$ denotes the partial likelihood given amino acid j at the descendent node d of the edge.

The update (A.2) can be obtained efficiently through a three-step update:

$$C_k = \sum_{j=1}^{c} U_{kj}L_j(d), \quad k=1,\ldots,c, \tag{A.3}$$

$$V_k = \exp[\lambda_k t]C_k, \quad k=1,\ldots,c, \tag{A.4}$$

$$L_i(p,e) = \sum_{k=1}^{c} U_{ik}^{-1}V_k, \quad i=1,\ldots,c. \tag{A.5}$$

Finally, the overall likelihood is obtained as $L=\sum_i \pi_i L_i(r)$, where $L_i(r)$ is the partial likelihood at the root.

Let A, M and E denote the CPU costs of addition, multiplication and calculation of $\exp(x)$. Counting operations over all values of k, the cost of (A.3) is $c(cM+(c-1)A)=c^2M+c(c-1)A$ and the cost of (A.4) is $2cM+cE$. The cost of (A.5), which needs to be repeated over all values of i, is the same as (A.3). Thus the total cost for an internal edge is $2c(c+1)M+2c(c-1)A+cE$. Efficient calculation of the external edge update (A.1) requires (A.4)-(A.5) but $C_k=U_{ks(d)}$ and so (A.3) can be ignored, giving cost $c(c+2)M+c(c-1)A+cE$. At each internal node, the products $L_i(p)=L_i(p,e_1)L_i(p,e_2)$ need to be calculated each with a cost of cM. Finally, the sum, $L=\sum_i \pi_i L_i(r)$ has a cost of $(c-1)A+cM$. With m taxa there are $m-3$ internal edges, m external edges and $m-2$ internal nodes, so the total cost is

$$\begin{aligned} C(PMSF) = {} & (2c(c+1)M+2c(c-1)A+cE)(m-3) \\ & +(c(c+2)M+c(c-1)A+cE)m+cM(m-2) \\ & +(c-1)A+cM \end{aligned} \tag{A.6}$$

For $LG+F+\Gamma$, a computational savings is possible because for each edge and k, $\exp[\lambda_k t]$ can be precomputed; by contrast PMSF repeats these calculations at each site. Thus

$$C(LG)=C(PMSF)-(2m-3)(cM+cE) \qquad \text{(A.7)}$$

A k-category mixture needs to repeat the likelihood evaluations for each component giving cost $kC(LG)$.

For the gcc C compiler, we approximated relative values of M, A and E using repeated multiplications, additions and exponentiations as $E=22A$ and $A=M$. Substituting these values into (A.6) and (A.7) gives $C(PMSF)/C(LG)$ decreasing from approximately 1.5 with 4 taxa to 1.4 as T gets large. We note that there are ways of reducing the cost of updating for $LG+F+\Gamma$ by precomputing substitution matrices. However, the calculations indicated here are more efficient overall because updates can be re-used when doing derivative calculations for edge lengths.

References

Aguinaldo A.M.A., Turbeville J.M., Linford L.S., Rivera M.C., Garey J.R., Raff R.A., Lake J.A. 1997. Evidence for a clade of nematodes, arthropods, and other moulting animals. Nature 387:489–493.

Akaike H. 1974. A new look at the statistical model identification. IEEE Trans. Automat. Control 19:716–723.

Barber, C.B., Dobkin, D.P., and Huhdanpaa, H.T., 1996. The Quickhull algorithm for convex hulls. ACM Trans. Math. Software 22:469–483, http://www.qhull.org.

Brinkmann H., van der Giezen M., Zhou Y., Poncelin de Raucourt G., Philippe H. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst. Biol. 54:743–757.

Brown M.W., Sharpe S.C., Silberman J.D., Heiss A.A., Lang B.F., Simpson A.G., Roger A.J. 2013. Phylogenomics demonstrates that breviate flagellates are related to opisthokonts and apusomonads. Proc. Biol. Sci. 280:20131755.

Daubin V., Gouy M., Perrière G. 2002. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. Genome Res. 12:1080–1090.

Delsuc F., Brinkmann H., Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. 6:361–375.

Drew B.T., Ruhfel B.R., Smith S.A., Moore M.J., Briggs B.G., Gitzendanner M.A., Soltis P.S., Soltis D.E. 2014. Another look at the root of the Angiosperms reveals a familiar tale. Syst. Biol. 63:368–382.

Goremykin V.V., Nikiforova S.V., Cavalieri D., Pindo D., Lockhart P. 2015. The root of flowering plants and total evidence. Syst. Biol. 64:879–891.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401–410.

Goldman N., Thorne J.L., Jones D.T. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics 149:445–458.

Guindon S., Dufayard J.F., Lefort V., Anisimova M., Hordijk W., Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59:307–21.

Halpern A.L., Bruno W.J. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol. Biol. Evol. 15:910–917.

Izquierdo-Carrasco F., Gagneur J., Stamatakis A. 2012. Trading running time for memory in phylogenetic likelihood computations. Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms. 86–95. Portugal: Science and Technology Publications.

Jones D.T., Taylor W.R., Thornton J.M. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8:275–282.

Kalbfleisch, J.G. (1985). Probability and statistical inference. Statistical inference, Vol. 2. New York: Springer.

Kocot K.M., Cannon J.T., Todt C., Citarella M.R., Kohn A.B., Meyer A., Santos S.R., Schander C., Moroz L.L., Lieb B., Halanych K.M. 2011. Phylogenomics reveals deep molluscan relationships. Nature 477:452–456.

Kuramae E.E., Robert V., Snel B., Weiss M., Boekhout T. 2006. Phylogenomics reveal a robust fungal tree of life. FEMS Yeast Res. 6:1213–1220.

Lanfear R., Calcott B., Ho S.Y.W., Guindon S. 2012. Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol. Biol. Evol. 29:1695–1701.

Lartillot N., Brinkmann H., Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol. Biol. 7(1 Suppl):S4.

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.

Lartillot N., Rodrigue N., Stubbs D., Richer J. 2013. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. Syst. Biol. 62:611–615.

Le S.Q., Dang C.C., Gascuel O. 2012. Modeling protein evolution with several amino acid replacement matrices depending on site rates. Mol. Biol. Evol. 29:2921–2936.

Le S.Q., Gascuel O. 2008. An improved general amino acid replacement matrix. Mol. Biol. Evol. 25:1307–1320.

Le S.Q., Gascuel O. 2010. Accounting for solvent accessibility and secondary structure in protein phylogenetics is clearly beneficial. Syst. Biol. 59:277–287.

Le S.Q., Gascuel O., Lartillot N. 2008a. Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics. 24:2317–2323.

Le S.Q., Lartillot N., Gascuel O. 2008b. Phylogenetic mixture models for proteins. Philos. Trans. Roy. Soc. London Ser. B 363:3965–3976.

Leebens-Mack J, Raubeson LA, Cui L, Kuehl JV, Fourcade MH, Chumley TW, Boore JL, Jansen RK, depamphilis CW. 2005. Identifying the basal angiosperm node in chloroplast genome phylogenies: sampling one's way out of the Felsenstein zone. Mol Biol Evol. 22:1948–1963.

Minh B.Q., Nguyen M.A.T., von Haeseler A. 2013. Ultrafast approximation for phylogenetic bootstrap. Mol. Biol. Evol. 30:1188–1195.

Neyman J., E.L. Scott. 1948. Consistent estimates based on partially consistent observations. Econometrica 16:1–32.

Nguyen L.-T., Schmidt H.A., von Haeseler A., Minh B.Q. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. Mol. Biol. Evol. 32:268–274.

Philippe H., Brinkmann H., Lavrov D.V., Littlewood D.T.J., Manuel M., Wörheide G., Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 9:e1000602.

Pisani D., Pettc W., Dohrmannd M., Feudae R., Rota-Stabellif O., Philippeg H., Lartillot N., Wörheide G. 2015. Genomic data do not support comb jellies as the sister group to all other animals. Proc. Natl. Acad. Sci. USA 112:15402–15407.

Pupko T., Huchon D., Cao Y., Okada N., Hasegawa M. 2002. Combining multiple data sets in a likelihood analysis: which models are the best? Mol. Biol. Evol. 19:2294–2307.

Rambaut A., Grassly N.C. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic Trees. Comput. Appl. Biosci. 13:235–238.

Raymann K., Brochier-Armanet C., Gribaldo S. 2015. The two-domain tree of life is linked to a new root for the Archaea. Proc. Natl. Acad. Sci. USA 112:6670–6675.

Robinson D. R., Foulds L. R. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.

Rodrigue N. 2013. On the statistical interpretation of site-specific variables in phylogeny-based substitution models. Genetics 193:557–564.

Sander C., Schneider R. 1994. The HSSP database of protein structure-sequence alignments. Nucleic Acids Res. 22:3597–3599.

Self S., Liang K. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions, J. Am. Stat. Assoc. 82:605–610.

Struck T.H., Paul C., Hill N., Hartmann S., Hösel C., Kube M., Lieb B., Meyer A., Tiedemann R., Purschke G., Bleidorn C. 2011. Phylogenomic analyses unravel annelid evolution. Nature 471:95–98.

Susko E., Field C., Blouin C., Roger A.J. 2003. Estimation of rates-across-sites distributions in phylogenetic substitution models. Syst. Biol. 52:594–603.

Susko E., Inagaki Y., Roger A.J. 2004. On inconsistency of the neighbor-joining, least squares, and minimum evolution estimation when substitution processes are incorrectly modeled. Mol. Biol. Evol. 21:1629–1642.

Telford M.J., Budd G.E., Philippe H. 2015. Phylogenomic insights into animal evolution. Curr. Biol. 25:R876–R887.

Wang H.C., Li L., Susko E., Roger A.J. 2008. A class frequency mixture model that adjusts for site specific amino acid frequencies and imporves inference of protein phylogeny. BMC Evol. Biol. 8:331.

Wang H.C., Susko E., Roger A.J. 2014. An amino acid substitution-selection model adjusts residue fitness to improve phylogenetic estimation. Mol. Biol. Evol. 31:779–792.

Whelan N.V., Halanych K.M. 2016. Who let the CAT out of the bag? accurately dealing with substitutional heterogeneity in phylogenomic analyses. Syst. Biol. doi: 10.1093/sysbio/syw084.

Whelan N.V., Kocot K.M., Moroz L.L., Halanych K.M. 2015. Error, signal, and the placement of Ctenophora sister to all other animals. Proc. Natl. Acad. Sci. USA 112:5773–5778.

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.

Wickett N.J., Mirarab S., Nguyen N., Warnow T., Carpenter E., Matasci N., Ayyampalayam S., Barker M., Burleigh J.G., Gitzendanner M.A., Ruhfel B.R., Wafula E., Der J.P., Graham S.W., Mathews S., Melkonian M., Soltis D.E., Soltis P.S., Miles N.W., Rothfels C.J., Pokorny L., Shaw A.J., DeGironimo L., Stevenson D.W., Surek B., Villarreal J.C., Roure B., Philippe H., dePamphilis C.W., Chen T., Deyholos M.K., Baucom R.S., Kutchan T.M., Augustin M.M., Wang J., Zhang Y., Tian Z., Yan Z., Wu X., Sun X., Wong G. K-S., Leebens-Mack J. 2014. A phylotranscriptomics analysis of the origin and diversification of land plants. Proc. Natl. Acad. Sci. USA 111:E4859–4868.

Yang Z. 1996. Maximum-Likelihood models for combined analyses of multiple sequence data. J. Mol. Evol. 42:587–96.