Are profile mixture models over-parameterized?

Hector Baños[1,2,3,*], Edward Susko[2,3], and Andrew J. Roger[1,3]

[1] *Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia, B3H 4R2, Canada*

[2] *Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, B3H 4R2, Canada*

[3] *Institute for Comparative Genomics and Evolutionary Bioinformatics, Dalhousie University, Halifax, Nova Scotia, B3H 4R2, Canada*

**hbanos@dal.ca*

Abstract

Site heterogeneity of the amino acid substitution process accounts for the biochemical constraints on the range of admissible amino acids at specific sites. Phylogenetic models of protein sequence evolution that do not account for site heterogeneity are more prone to long-branch attraction artifacts.

Profile mixture models are used to model site heterogeneity. Even though model, tree, and mixing parameters are statistically consistent, the performance of these models with short alignments is unclear. Here we explore the behavior of tree topology estimates and marginal cumulative distributions with short simulated alignments. We find that over-parameterization is not a problem for complex profile mixture models and that simple models behave poorly. Misspecification of the frequency distributions does not cause a problem if the estimated cumulative distribution function adequately approximates the true one. Also, we find that misspecification of the exchangeabilities can severely affect parameter estimation and that an increase in likelihood does not necessarily reflect better tree estimation. Although the inclusion of more taxa often helps, it can hurt estimation if the exchangeabilities are badly misspecified.

Finally, we explore the effects of including an 'F-class' with the overall amino acid

frequencies of the dataset as an additional class in the profile mixture model. Surprisingly, the F-class does not seem to help parameter estimation significantly, and it can decrease the probability of correct tree estimation, depending on the scenario, despite the fact that it tends to improve likelihood scores. We also investigate this with several empirical data sets.

Key words: Phylogenetics; Mixture model; Frequency profile mixtures; Long-branch attraction.

Phylogenetic methods have been used to resolve many deep phylogenetic problems in the tree of life (Brown et al. (2013); Daubin (2002); Pisani et al. (2015); Raymann et al. (2015); Wickett et al. (2014)). To decrease estimate variability, these models require a large number of orthologous genes. The alignments of multiple genes (proteins) are either concatenated into a 'supermatrix' from which trees are estimated or individual gene/protein trees are first estimated and then combined using supertree or 'species tree' methods. In either case, as more genes or proteins are considered, systematic biases can arise (Philippe et al. (2011)), underscoring the importance of adequately modeling the nucleotide or amino acid substitution process.

The substitution process of amino acid sequences is usually modeled as a site-independent Markov process in a tree. The most common approach assumes constant stationary frequencies of the amino acids and a constant matrix of exchangeabilities throughout the tree. The amino acid frequencies are usually estimated from the observed frequencies in the entire alignment. The matrix of exchangeabilities is fixed *a priori*, chosen from a set of empirically defined matrices, see for example Jones et al. (1992); Le and Gascuel (2008a); Whelan and Goldman (2001). Also, it is customary to consider different rates across sites to accommodate faster or slower substitution processes at different sites (see for example Yang (1994)). These models with almost the same

substitution process at each site and where the only difference comes from distinct rates across sites, are known as *site frequency homogeneous models*. From now on, and as it is customary, we refer to these as *site-homogeneous*, although these are not strictly homogeneous since site-rates may differ per site.

However, there are different ranges of amino acids admissible at sites in proteins because of functional or structural restrictions (Franzosa and Xia (2009); Goldstein (2008); Pál et al. (2006)). These ranges can vary widely, from a few or just one, to essentially all possible amino acids at a site (Halpern and Bruno (1998); Lartillot et al. (2007); Lartillot and Philippe (2004); Wang et al. (2008)). Consequently, site-homogeneous models, which overlook this across-site amino acid frequency heterogeneity, are less biologically plausible and are prone to long-branch attraction (LBA) artifacts (Feuda et al. (2017); Lartillot et al. (2007); Simion et al. (2017); Wang et al. (2008); Williams et al. (2013)). LBA is a pervasive systematic bias in tree estimation whereby distantly-related groups with long branches are artefactually grouped together (Felsenstein (1978); Philippe and Laurent (1998)). Partition (Lanfear et al. (2016); Pupko et al. (2002); Yang (1996)) and mixture models (Lartillot and Philippe (2004); Le and Gascuel (2008a); Schrempf et al. (2020); Le and Gascuel (2008b); Wang et al. (2008)) have been used to model heterogeneity of the amino acid substitution process across sites.

The CAT model (Lartillot and Philippe (2004)), a popular Bayesian mixture model, was shown to be less prone to LBA artifacts and to fit data better than site-homogeneous models (Lartillot et al. (2007)). In this model, all frequency vectors are assumed to be independently and identically drawn from a Dirichlet process model which effectively allows non-parametric estimation of the mixing distribution. Unfortunately, in its current implementation (Lartillot et al. (2013)), convergence may not be achieved in practice for large data sets. In the maximum likelihood framework, models accounting for site heterogeneity include mixture of the substitution rate matrices predefined for sites coming from different secondary structural elements and surface accessibility classes (Goldman

et al. (1998); Le and Gascuel (2008a, 2010)), or for different site rates (Le et al. (2012)), and a mixture of amino acid site frequency profiles (Schrempf et al. (2020); Le and Gascuel (2008b); Wang et al. (2008, 2014)). The latter, known as *profile mixture models*, have become widely used for analyses of deep phylogenetic problems.

Frequency vectors and weights pre-estimated from data bases of alignments are frequently used to reduce the complexity and computational cost of estimation with profile mixture models (Schrempf et al. (2020); Le and Gascuel (2008b); Wang et al. (2008)). We refer to a set of frequency vectors with their corresponding weights as a *mixing distribution*, and we refer to the indices of the frequency vectors in the mixing distribution as the *classes*. Similar to the empirical estimates of rate matrices, such mixing distributions are estimated from large data sets such as those described in Dufayard et al. (2005) and Sander and Schneider (1994). The techniques used to obtain empirical estimates of these mixing distributions vary. For example, in Le and Gascuel (2008b) the authors introduced six mixing distributions having 10, 20, 30, 40, 50, and 60 classes that were estimated from large data sets by ML estimation. These are known as the C10-C60 (or generically CXX) mixing distributions. Schrempf and colleagues (Schrempf et al. (2020)) used K-means and the CAT model to estimate empirical mixing distributions ranging from 4, 8, 16, up to 4096 classes. These are known as the UDM mixing distributions.

Profile mixture models are less susceptible to LBA than site-homogeneous models (Wang et al. (2008)). Also, as carefully described in the next section, these models have desirable properties when inferring parameters. For example, identifiability of the tree and mixing distribution is known to hold for a large subclass of models (Yourdkhani et al. (2021)). As a consequence, the tree and the mixing parameters are statistically consistent even with a large number of profiles. Informally speaking, this means that, if the model is correctly specified, one can effectively estimate the true parameters as the number of sites increases.

Although the identifiability and consistency properties satisfied by profile mixture

models are desirable in any modeling context, there is no guarantee that such models will have good small sample properties. Nor it is known what problems may arise from model misspecification. Thus there is a need to explore the performance of profile mixture models through simulations of short alignments both with and without model misspecification. One of our main motivations is to determine whether the large numbers of parameters in profile mixtures create problems with small samples. Specifically, we want to determine if there is excessive variability in estimates from models with too many parameters relative to the sample size.

By varying several parameters of empirically-derived profile mixture models, we simulated distinct alignments of lengths 300, 600, and 1000 (the approximate lengths of true alignments of single proteins). A detailed explanation of the different simulation model settings is provided below. We fit distinct mixing distributions and matrices of exchangeabilities to each simulation. We assessed model performance using four criteria described in detail in the Materials and Methods. Two of these criteria concern the tree topology MLE accuracy and variability. The other two are a measure for comparing the marginal cumulative distribution functions (CDFs) inherited from the observed and expected mixing distributions. These CDFs, properly described in the following section, are an alternative way to re-parameterize profile mixture models.

All findings are presented later in detail, but the major highlights include the following:

- When the exchangeabilities and frequency vectors are correctly specified, there is no evidence of model over-parameterization (over-fitting), even when there are many classes with zero weight estimates. This relates to a concern articulated in several studies (e.g. Anderson and Lindgren (2021); Li et al. (2021)) that fitted complex mixture models with classes estimated to have zero weights are over-parameterized and should be avoided in favor of simpler models. Also, the inclusion of more taxa improves tree estimation.

- When there is misspecification of the frequency classes, we observe that tree estimation is not necessarily acutely affected. If the set of frequency vectors is sufficiently rich, the estimated CDF closely approximates the CDF of the generating model, and when that occurs, the frequency of correct tree estimations is large.

- Severe problems can arise from the misspecification of the exchangeabilities. We observe that this scenario can lead to a bias in the MLE mixture weights. This bias favors parameters that maximize the likelihood but decrease the similarity between observed and expected CDF. This produces a decline in tree estimation accuracy. Under these conditions, adding taxa does not necessarily improve tree estimation.

We also explore the effects of the "F-class," a class that is defined from the empirical frequencies of amino acids from the overall alignment, that is often included as an additional class in models to account for remaining sites in the data that are not well modeled by the fixed empirically-derived frequency vectors. However, we find that the F-class does not significantly improve tree estimation, and, in some cases, may compromise accuracy. This exploration is complemented by looking at empirical data. Our analyses suggest that while the F-class increases the likelihood significantly, it may lead to erroneous tree estimation.

Materials and Methods

Mixture models and over-parameterization with large samples

In this section, we define and elaborate on some theoretical properties of profile mixture models. Re-parameterizing the mixing distributions as CDFs provides insight into why over-parameterization is less of a problem for profile mixture models than it might be for models without stringent parameter constraints. Also, we briefly discuss known identifiability results for such models.

Roughly speaking, profile mixture models are mixtures of time-reversible models,

with a common exchangeability matrix R. The parameter space Θ of a profile mixture model with C classes is defined by:

(i) A rooted metric tree T on N taxa.

(ii) A symmetric 20×20 matrix of non-negative exchangeabilities R.

(iii) For $c = 1, 2, ..., C$, a frequency distribution vector $\boldsymbol{\pi}_c$, and a weight w_c, with $w_c > 0$ and $\sum_{c=1}^{C} w_c = 1$.

(iv) A collection of K scalar rate parameters $\{r_k\}$, with $r_k \geqslant 0$, and rate weight d_k, with $d_k > 0$ and $\sum_{k=1}^{K} d_k = 1$.

The substitution process of a profile mixture model is as follows: for each site, a frequency vector $\boldsymbol{\pi}_c$ is sampled with probability w_c; and a rate parameter r_k is sampled with probability d_k. Evolution of a sequence at a site is then according to a continuous Markov substitution process over tree T with exchangeabilities R, root distribution $\boldsymbol{\pi}_c$, and rate r_k. For a given site pattern $\boldsymbol{x}_i$, the likelihood function is determined by a weighted average of partial site likelihoods conditional on each site-profile class and site-rate class:

$$L(\theta|\boldsymbol{x}_i) = \sum_{c=1}^{C} w_c \sum_{k=1}^{K} d_k P(\boldsymbol{x}_i|T, R, \boldsymbol{\pi}_c, r_k),$$

where $\theta = (T, R, \{\boldsymbol{\pi}_c\}, \{w_c\}, \{r_k\}, \{d_k\}) \in \Theta$.

For a model with fixed frequency variables, a natural way of parameterizing the mixture model is in terms of its weights, w_c, $c = 1, \ldots, C$. This leads to models of differing dimensions C that, as mentioned before, can get very large, raising concerns about over-parameterization.

An alternative way of parameterizing the mixture is in terms of its cumulative distribution function (CDF):

$$G(\boldsymbol{\pi}) = P(\boldsymbol{\Pi} \leqslant \boldsymbol{\pi}),$$

where $\boldsymbol{\Pi}$ represents the random frequency vector for a site. This allows one to express mixing distributions with differing components ($C = 20$ say or $C = 60$) as being in the

same parameter space. However, now the frequency mixture parameter space, which is a space of distribution functions, is infinite-dimensional, and would appear an extreme case of over-parameterization.

Surprisingly, even for this infinite dimensional space of distribution functions, estimation of both the mixing distribution and structural parameters like the tree is frequently still consistent as was shown by Kiefer and Wolfowitz (1956) for a wide class of models under mild regularity conditions. Most of these conditions are expected to hold for the models considered here. A more detailed explanation of how the results in Kiefer and Wolfowitz (1956) apply to our context is given in the Appendix.

The implication of Kiefer and Wolfowitz (1956) is that class frequency mixture models are not overparameterized, at least with large samples. The reason for this is that the space of all distribution functions as a space of functions is relatively "small" in the mathematical sense of being a compact space (a closed and bounded space in our setting). An alternative way of seeing why this is the case is to note that the w_c are restricted to be non-negative and sum to one. By contrast in cases of true over-parameterization, parameters are unrestricted (for example, a regression model where there are more predictors than observations).

Another surprising result of estimation within the mixing distribution setting is that even if parameter estimation is unrestricted and any mixing distribution is allowed, the maximum likelihood estimator will be a finite mixing distribution: i.e. it will be describable in terms of a fixed set of weights $w_1, \ldots, w_C$ for some C. This is an implication of the results of Lindsay (1983) as detailed in the Appendix.

Adding to this, in Yourdkhani et al. (2021) identifiability results are reported for a large family of profile mixture models. These authors showed generic identifiability of the tree and mixing parameters for models with $C \cdot K < 72$, where C is the number of classes and K the number of rates, trees with more than 8 taxa, and where no parameters in Θ are assumed to be fixed. As discussed in the Appendix, these results also apply to some of

the models with fixed frequency vectors considered here. We conjecture that, generically, identifiability of the tree topology can be achieved for models with fixed frequency vectors with C classes, K rates and m taxa, for some $C \cdot K > 72$, and all $m > 8$. This will be explored in future work.

Simulation setting

In this section, we describe all the different parameters used to simulate alignments under profile mixture models. These are presented below in the following order: (1) the trees; (2) mixing distributions; (3) exchangeability matrices; (4) sequence lengths; and (5) rate parameters.

By choosing different combinations of parameters, we simulated a total of 108 scenarios. For each scenario, 100 simulations were performed using `Alisim` (Ly-Trong et al. (2021)). We now proceed to describe all the choices of parameters.

Trees Nine different trees are considered for these simulations. All trees have the 'structure' of tree T shown in Figure 1. The features that vary per tree are the length of a single edge l, where $l \in \{0.005, 0.02, 0.05\}$ and the number of taxa at each polytomy m, where $m \in \{1, 2, 3\}$. We denote each of the trees by $T_{6m}(l)$.

The structure of T is chosen since it is often a tree susceptible to LBA artifacts. For fixed l and changing m, the simulations are, in effect, all from the same tree but with differing levels of taxonomic sampling from the 6 clades. By increasing the number of taxa, we obtain more information on the frequency vectors. By decreasing the edge length l, we make the tree more susceptible to LBA artifacts whereby the f-clade (i.e. clade including taxa $f_1, f_2, ..., f_m$) and the e-clade (i.e. clade with taxa $e_1, e_2, ..., e_m$) group together to the exclusion of the other clades.

Mixing distributions For each tree, we simulated data using ten different mixing distributions. One of these is the model C60 as defined in Le and Gascuel (2008b), which

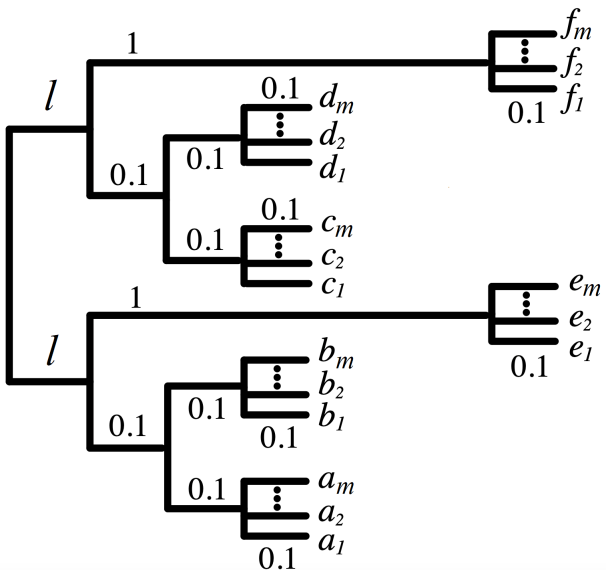

Fig. 1. The main structure of the tree where all the simulations were conducted. Only the edge lengths l and the number of taxa at each polytomy m are variable. This tree has $6m$ taxa

has 60 classes. Another seven are built on the frequency vectors of C60. Specifically, for $i \in \{10, 15, 20, 30, 40, 50, 60\}$, we defined the mixing distribution C60[i] by choosing i frequency vectors from C60 at random and assigning them non-zero weights sampled from a Dirichlet distribution with concentration parameters $\boldsymbol{\alpha} = \mathbf{1}$. We denote by $\mathfrak{C}60$ the set of mixing distributions C60[i], for all i, including C60. The two remaining mixing distributions used to simulate data are known as UDM-0256, and UDM-4096 (Schrempf et al. (2020)), that have 256 and 4096 classes respectively. Specifically, we used the non-transformed mixing distributions denoted as UDM-0256-None and UDM-4096-None in Schrempf et al. (2020).

In reality, every site in a protein has a unique physicochemical environment that would likely be better fit by a site-specific frequency vector. Therefore profile mixture models with fixed frequencies try to approximate commonly occurring patterns amongst sites represented by site classes. By simulating under C60, the most complex of the CXX mixing distributions, and the complex UDM mixing distributions, we try to emulate real data. The C60[i] distributions reflect the scenario wherein, for a small sample, not all

relevant frequency vectors are represented, nor are distributed as in C60. We note that the UDM mixing distributions include additional components that are not in the CXX distributions but because the data-sets used to estimate both CXX and UDM mixing distributions overlap, so some similarities in their frequency profiles are expected.

Exchangeability Matrices, Sequence Lengths, and Site Rate Variation For all combinations of mixing distributions and trees, we simulated data using the exchangeability matrix from the LG model (Le and Gascuel (2008a)); we refer to this as the LG matrix. We also used a "POISSON" exchangeability matrix, a matrix with equal exchangeabilities, but only for simulations involving the two UDM mixing distributions.

For all combinations of mixing distributions, trees, and matrices, we used sequence lengths of 300, 600, and 1000 amino acids. In practice, alignments of length 300 are more typical for single protein data sets. The other choices of sequence lengths are meant to capture the effects in tree estimation with increasing sequence length.

Lastly, for all simulations we used four rate parameters coming from a discrete-$\Gamma(4)$ distribution (Yang (1994)) with $\alpha = 0.5$.

Fitted models and Precision of parameter estimation

In this section, we describe how different choices of fixed frequency vectors and exchangeabilities were fitted to the simulations. We also introduce four criteria used to evaluate model fitness.

When fitting a model to data, estimated parameters were obtained by maximum likelihood. IQ-tree2 (Minh et al. (2020)) was used to get the maximum log-likelihoods and the estimators (MLEs) for all simulations. The parameters that were optimized by IQ-tree2 are the tree topology, edge lengths, and weights of the frequency vectors. All other parameters including the frequency vectors, the matrix of exchangeabilities, and the rate parameters, were supplied as fixed values to IQ-tree2.

The LG matrix was the only matrix used to simulate data under the mixing distributions in $\mathfrak{C}60$. To these simulations, we fit the LG matrix, the F-class, and various frequency vectors. Specifically, the frequency vectors fitted included C60, C40, C30, C20, defined in Le and Gascuel (2008b), LG (Le and Gascuel (2008a)), LG4X (Le et al. (2012)), CK36 as defined below, as well as the frequency vectors in $\mathfrak{C}60$ used to simulate the data. In these analyses, we explored mainly two things: (1) Possible model over-parameterization from fitting more general models to short alignments; and (2) misspecification of the frequency vectors by using models that have different frequency vectors than those of the generating model.

The frequency vectors CK36 were obtained from the cluster centers derived from a k-means algorithm (Hartigan and Wong (1979)) on all the classes of C20, C30, C40, and C60 with $k = 36$ (the choice of k was determined by the elbow method (Thorndike (1953))).

For data generated under the UDM mixing distributions, we fitted the exchangeability matrices POISSON and LG, with and without the F-class, and frequency vectors of C60, C40, C30, C20, CK36, LG4X. We also fitted the POISSON model for data generated under POISSON exchangeabilities and the LG model for data generated under LG exchangeabilities. In these analyses we primarily explored the effect of: (1) Misspecification of the frequency vectors; (2) misspecification of the matrix of exchangeabilities; and (3) use of the F-class in estimation. Recall that only the mixing weights, the tree topology, and edge-lengths are optimized, everything else is fixed before the maximization of the likelihood.

To simplify the presentation of results, we only considered a subspace of tree space. In preliminary results, we computed the likelihoods of all 105 6-taxon unrooted topologies for data generated under $T_6(0.005)$. We noted that there were two tiers in terms of log-likelihood values; these occurred regardless of the generating and fitted models. One tier consists of the log-likelihoods of the 35 topologies shown in Table S1 in the

Supplementary Material, and the other consists of the remaining 70 trees. The first tier showed significantly larger log-likelihood values, see Figure S1 in the Supplementary Material. The 35 topologies in the first tier are all the topologies displaying the embedded quartet tree $AB|CD$. This shows, as expected, there are no problems with estimating the relationships amongst the taxa in this quartet. The main difficulty was instead determining the correct placement of the long branches because of the LBA-related artifacts.

Therefore, for simplicity, in all cases we restrict the tree space considered to just these 35 tree topologies by substituting $X \in \{A, B, C, D, E, F\}$ by the adequate m-taxon polytomy. We also consider in this tree space the 'star tree' topology obtained from T in Figure 1 by setting $l = 0$.

We now proceed to introduce the criteria used to compare the overall model performance.

Mean Integrated Squared Error and Maximal Difference As mentioned earlier, profile mixture models can be parameterized as CDFs. Therefore, a reasonable way to measure the precision of parameter estimation is comparing how closely the observed CDF resembles the true one. Unfortunately, assessing the precision of parameter estimate *via* the expected and observed CDFs is burdensome due to the complexity of the 20-dimensional space in which they reside. Here we use marginal CDFs instead to assess the precision of parameter estimation.

To compare marginal CDFs we used two measures. The first one is the *Mean Integrated Squared error* (MISE) (Scott (1992)), also known as L^2 risk function, which is defined as follows

$$\text{MISE} = \frac{1}{20}\sum_{i=0}^{20}\int_0^1 (G_i(x) - \hat{G}_i(x))^2 dx,$$

where G_i is the true marginal CDF corresponding to amino acid i, and $\hat{G}_i(x)$ is the marginal CDF obtained from the estimated mixing distribution. Here all the integrals were computed using the function `integrate` from the R package `pracma` with default settings.

The second measure consists in the maximum difference (MD) between estimated marginal CDFs and the true ones. This is borrowed from the Kolmogorov–Smirnov test (Massey (1951)), and it is defined as follows:

$$\text{MD} = \frac{1}{20}\sum_{i=0}^{20} \max_{x\in[0,1]}\{|G_i(x) - \hat{G}_i(x)|\},$$

where G_i and $\hat{G}_i(x)$ are as defined for the MISE. Here the maximum of the absolute difference is computed using the function `optimize` from the R package `stats` with default settings.

For a given choice of parameters, we report the mean MISE and MD for all 100 simulations. Note that in both cases, as these measures approach to zero, the observed CDF approaches the true CDF. While these two measures cannot be used in practice since the true CDF is then unknown, in this case and as detailed in the Results section, these measures help us assess model over-parameterization among other things.

In order to assess the precision of parameter estimation in a more standard way, we also looked at two more criteria based on the tree topology estimate.

Overall Accuracy and Proportional Mode The following criteria, *overall accuracy* denoted OA, is a standard way to evaluate the precision of parameter estimation. Given a set of simulations, OA is defined as the proportion of these where the true tree topology was the one maximizing the likelihood.

The next criterion, the *proportion of settings where the true tree was the mode of the distribution of estimated trees*, denoted PM, evaluates the precision of parameter estimation in a manner that can indicate whether there are biases in estimation. PM consists of the proportion of settings where the true topology was chosen the most. Given several sets of simulations, PM is the proportion of these sets where the true tree topology was the MLE more often than any other topology. For a given method, the tree estimated most frequently is the mode of the distribution of estimated trees. So alternatively, PM is the frequency with which the true tree is the mode of the distribution of estimated trees.

To account for some sampling variability in PM, in each scenario we tested if the proportion of times the true topology was chosen was significantly higher than the runner-up topology using a binomial test. PM is obtained from dividing the number of scenarios where the true topology is significantly more likely than the runner-up (rejecting the null hypothesis $H_0 : p = 0.5$) by the total number of scenarios.

Note that OA and PM may not necessarily be strongly correlated. Suppose, for instance, that for a given model there is a bias in estimation towards certain trees under certain settings. That could result in a large OA because of the large frequency of estimations of the true tree in settings where the model is biased towards it. But if the true tree is not always favored, the true tree would not be most frequently estimated, leading to a small PM. On the other hand, one could have really low OA but no other tree is chosen more times, leading to a high PM. Ideally, one would like to see both high OA and PM, which would indicate that the true tree is being chosen the most and with little variability across scenarios.

Results

We present the results according to the set of mixing distributions we used to simulate the data. We start with $\mathfrak{C}60$, then UDM, followed by the empirical data-sets. For the remainder of the text and for simplicity, we denote by C60L the model C60 as defined in Le and Gascuel (2008b) including both frequency classes and optimal weights from that study. Then, when we discuss fitting one of the CXX models we are referring strictly to fitting just its frequency classes.

Simulated Data Under $\mathfrak{C}60$ *Mixing Distributions*

Table 1 displays the mean MISE for data generated under 72 different simulation conditions total; i.e. under eight mixing distributions in $\mathfrak{C}60$ and nine trees. The eight mixing distributions included C60L and C60[i] for i in $\{10, 15, 20, 30, 40, 50, 60\}$. We see in

	PF+F	**C60+F**	**C40+F**	**CK36+F**	**C30+F**	**C20+F**
300	0.00037	0.00042	0.00154	0.00118	0.00199	0.00333
600	0.00020	0.00023	0.00137	0.00102	0.00179	0.00309
1000	0.00013	0.00015	0.00129	0.00095	0.00170	0.00299

Table 1. The mean MISE for data generated under distributions in $\mathfrak{C}60$. For each of the 72 scenarios (9 trees and 8 models) per sequence length, we compute the mean MISE for the 100 simulations. Then we take the mean of all these values, which are the entries in this table. Label PF+F represents the overall performance in MISE when fitting the generating classes per scenario with the F-class.

this table, as expected, that estimating models that included only the frequency classes used to generate the data plus an F-class, denoted here as perfect fit (PF)+F, are those with the lowest MISE. The second lowest value is achieved by C60+F model, followed by CK36+F, C40+F, C30+F, and C20+F, respectively. In this case, we note that the MISE of PF+F and C60+F are really close. This behavior is similar for the MD criterion, as seen in Table S2 in the Supplementary Material. These observations suggest that even though C60+F fits 61 classes to the data, it still has a much better fit than the CXX+F models with fewer classes even though, for many of the simulation settings, there were many fewer classes present. This demonstrates that over-parameterization is not a problem for complex, correctly specified models, even for models having several classes with zero weights. Table S3 in the Supplementary Material shows the mean normalized MISE for the same data as in Table 1, where MISE for any given scenario was re-scaled to give a sum of 1 over all fitted models.

This shows that the mean MISE values in Table 1 adequately consolidate all scenarios and no biases between classes are introduced by a scenario with considerably larger MISE values.

Figure 2 shows the plots of average OA (A) and PM (B) over all data generated under mixing distributions in $\mathfrak{C}60$. Each dot in OA represents the proportion of times the tree was correctly inferred over 2400 simulations (3 values of l and 8 sets of classes in $\mathfrak{C}60$, with 100 repetitions each), and PM is a proportion over 24 distinct scenarios. For alignments of length 300, fitting C60+F, C40+F, C30+F, C20+F, or CK36+F, produces,

on average, no significant difference in OA. However, C60+F and CK36+F have better PM values. On the other hand, models LG+F and LG4X+F perform poorly as reflected by both OA and PM.

For longer alignments, C20+F has a significantly lower OA than C60+F, C40+F, C30+F, and CK36+F. This shows how as sample size increases, more complex models that approximate the true CDF better have superior performance. In this case, fitting C60+F and CK36+F still yields the best PM values. Overall these results reinforce the inference that over-parameterization of C60+F does not cause problems and misspecification of the frequency vectors (e.g. for CK36+F) does not compromise tree estimation if the estimated CDF adequately approximates the true CDF.

One concern is that C60+F could be fitting better on average because it closely resembles two of the mixing distributions – C60L and C60[60] – used in the foregoing simulations. However, C60+F also behaves well for data generated with fewer classes. Table S4 in the Supplementary Material shows the OA for data generated under C60[i], with $i \in 10, 30, 60$, $T_6(0.005)$, and different sequence lengths. This table shows that there is little evidence of over-fitting when the estimating model has many more classes than the generating model. This is also reflected in the MISE scores when considering different generating classes (Supplementary Material Table S5).

Figure 2 also shows that increasing numbers of taxa improves tree estimation. When the number of taxa increases, MISE and MD scores decrease/improve (Table 2) and this behavior is consistent across models. As expected, we also observe that tree estimation accuracy also improves as sequence length increases. Analogously, MISE and MD scores decrease as sequence length increases (Table 2).

Simulated Data Under UDM Mixing Distributions

Table 3 displays the mean MISE values over data generated under UDM-0256 with POISSON exchangeabilities. The MISE values are considered separately when fitting CXX

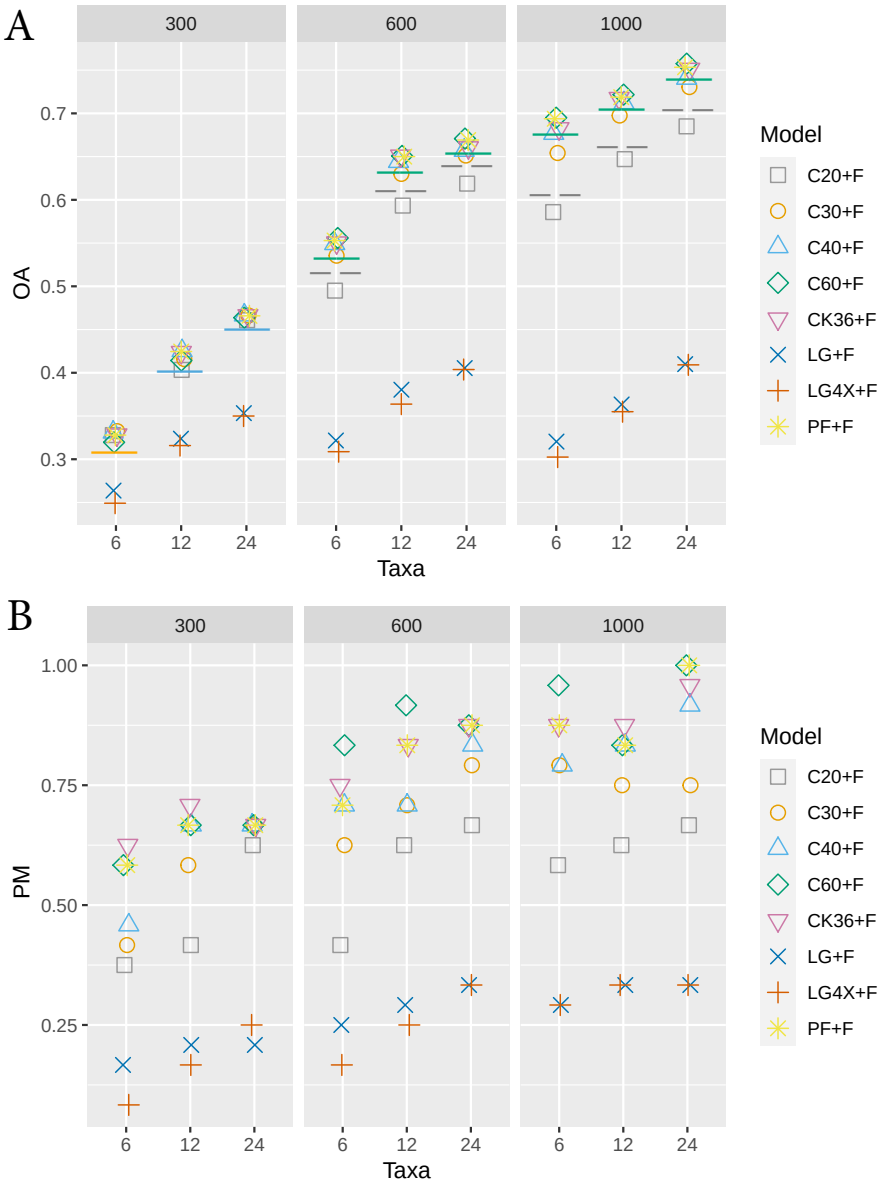

Fig. 2. (A) the plot of the OA values for various models fitting data generated under mixing distributions in $\mathfrak{C}60$ fitted to all frequency vectors. Label PF denotes the frequency vectors used to generate the data. The x-axis represents the number of taxa on the tree. The plot is divided by sequence lengths of 300, 600, and 1000. The lower bound of the 95% confidence interval (CI) of the model with the highest OA is depicted with a solid line whose color is in agreement with such model. Depicted with a dashed gray line is the higher bound of the CI of classes in C20 in the cases where such model was significantly worse than the best model. (B) A similar plot to that on top but for PM. For this case, no confidence interval can be computed.

	MISE			MD		
Taxa	**300**	**600**	**1000**	**300**	**600**	**1000**
6	0.00040	0.00020	0.00012	0.178	0.136	0.111
12	0.00031	0.00016	0.00008	0.155	0.117	0.085
24	0.00024	0.00013	0.00007	0.127	0.097	0.071

Table 2. The MISE and MD for fitted C60+F to data generated under $T_{6m}(0.02)$ and C60[15] for all m. One can see a decrease in MISE and MD by either increasing the number of taxa, or the sequence length.

	LG matrix			POISSON matrix		
	300	**600**	**1000**	**300**	**600**	**1000**
C60+F	0.00724	0.00783	0.00822	0.00083	0.00064	0.00055
C40+F	0.00767	0.00822	0.00861	0.00091	0.00071	0.00062
CK36+F	0.00807	0.00856	0.00890	0.00089	0.00071	0.00063
C30+F	0.00852	0.00901	0.00931	0.00096	0.00076	0.00068
C20+F	0.00919	0.00964	0.00990	0.00111	0.00092	0.00085
C60	0.00357	0.00378	0.00388	0.00080	0.00062	0.00054
C40	0.00333	0.00337	0.00343	0.00087	0.00069	0.00061
CK36	0.00435	0.00448	0.00457	0.00086	0.00069	0.00062
C30	0.00422	0.00438	0.00444	0.00090	0.00073	0.00065
C20	0.00398	0.00399	0.00405	0.00103	0.00088	0.00082

Table 3. The mean MISE for data generated under UDM-0256 and POISSON exchangeabilities. For each of the 9 scenarios (9 trees) per sequence length, we compute the mean MISE for the 100 simulations. This is done separately when fitting the LG and POISSON matrices.

models and CK36 (with and without the F-class) to the POISSON and LG matrices. In the former case, there is misspecification just of the classes whereas, for the latter, both the classes and exchangeabilities are misspecified.

When fitting the LG matrix, i.e where there is misspecification of the exchangeabilities, we see a significant difference in model fit between models including the F-class and omitting it. The MISE scores are elevated for models including the F-class relative to those without and this effect is independent of the sequence length (Table 3). We believe this is not directly an artifact of the F-class *per se*, as it is discussed below.

When fitting the correctly specified POISSON matrix, for any given sequence length and set of fitted frequency vectors, the mean MISE is comparable whether the F-class is included or omitted. In most cases, we see a slightly better MISE when excluding the F-class although the difference is minuscule and most likely not significant. We note

that the classes in C60 have the best MISE scores across all sequence lengths, but many of the other CXX models also performed well; C20 yielded the poorest scores overall. In this same table, we also see how MISE decreased as sequence length increased when there is no misspecification but this trend did not necessarily hold when there was. We believe this comes from the fact that as more data becomes available, more classes can be misestimated introducing more error.

Figures 3 and 4 show the plots of average OA (A) and PM (B) for data generated under UDM distributions and POISSON exchangeabilities. The former figure shows data fitted using models with POISSON exchangeabilities, and the latter, LG exchangeabilities. In these plots, each dot in the OA plot represents the proportion of times the tree was correctly inferred over 600 simulations (3 values of l, 2 sets of UDM classes, and 100 repetitions), and for PM the proportion over 6 distinct scenarios.

Figure 3 shows the case when there was misspecification of classes only. Recall that the mixing distributions in UDM have 256 and 4096 frequency vectors, thus there is significant misspecification when fitting CXX and the CK36 models. Nevertheless, we saw that in all cases, fitting with any of the CXX mixtures led to reasonable performance. The OA estimates were even close to the case of no misspecification of the frequency classes in Figure 2.

In this case, there seems to be no significant difference in OA between fitting the F-class or discarding it. However, fitting without the F-class yielded, in some cases, better PM values. Similar to the $\mathfrak{C}$60 case, fitting with LG4X and POISSON with no site-profile mixture model also behaved both very poorly. In this case, we also see how the increase of taxa monotonically improved tree estimation. The exception was for sequence lengths of 600 and 1000 when increasing taxa from 12 to 24. We consider these to reflect a tie in performance accounting for variability due to the finite number of simulations. Similar conclusions can be drawn for the case where the only difference was that the data was generated and fitted using the LG matrix instead of POISSON (Supplementary Material

Figure S2).

When there is misspecification of both classes and exchangeabilities as shown in Figure 4, we observe that for all cases, except one, OA was equal or significantly worse compared to Figure 3 and Figure S2 in the Supplementary material. We believe this is similar to the behavior shown in Table 3 described above. Furthermore, in some cases, an increase in numbers of taxa led to a decrease in OA (see Figure 4, 1000 sites).

The most striking, and perhaps surprising, feature in Figure 4 is that fitting with the F-class often led to a substantial drop in OA compared to when it is omitted. To investigate this further, we explored the impact of the F-class on the likelihood scores and mixing weights of the fitted models when exchangeabilities were misspecified (LG) versus correctly specified (POISSON). These analyses were based on 27000 observations (9 trees, 3 sequence lengths, 2 UDM distributions, 5 fitted models, 100 simulations) and the results are shown in Figure 5. When there is misspecification of the exchangeabilities, the F-class frequently improves the likelihood values substantially, whereas when the exchangeabilities are correctly specified, only modest increases in likelihood are seen (see Figure 5 (A)). The frequently large increases in likelihood values with the F-class in the case of the LG exchangeabilities is surprising because, as indicated in Figure 4 and Table 3, better OA estimates and lower MISE scores are obtained in this case when there is no F-class.

Since models without the F-class are special cases of the comparable models that include F-classes, we should always expect an increase in likelihood when fitting the latter models. When there is no misspecification of the frequency classes, a crude approximation is that the null distribution has a mean of 5 and a standard deviation of 5.48 (from a mixture of a degenerate uniform[0] and a χ^2 with 20 degrees of freedom, see Self and Liang (1987)). When there is no misspecification, simulations have a mean of 3.8 and a standard deviation of 3.4, smaller than the natural increase in likelihood, alluded to above, that are expected with increases in the number of parameters estimated. By contrast, with model misspecification, the mean and standard deviation are 34.9 and 19.7, respectively. Thus

when there is no misspecification of the exchangeabilities, differences in log-likelihood are pretty small and would be judged small relative to crude chi-square approximations. By contrast, with misspecification, such differences are very large compared to expectations based on the number of parameters estimated.

We also investigated the impact of misspecification of exchangeabilities on the estimated weight of the F-class (Figure 5B). Figure 5 (B) contains two overlapping plots. When exchangeabilities are correctly specified, the F-class weights tend to be relatively small (e.g. mean weight of non-misspecified = 0.11), whereas when they are misspecified the weight distribution shifts dramatically to adopt larger values, often exceeding 0.5 (e.g. mean weight of misspecified = 0.62). Clearly, the weights of the F-class are far from zero in the latter case. We explore this bias further below in relation to the uniformity of frequencies at sites as measured by Shannon entropy.

The Shannon entropy, as defined in our context

$$H(\boldsymbol{\pi}) = -\sum_{j=1}^{20} \pi_j \ln(\pi_j),$$

is a common measure of the degree of uniformity of the amino acid frequencies at sites. We note that when there is misspecification of the exchangeabilities, there is a bias towards frequency classes with high entropy. In more than 80% of the 27000 data sets where we fitted a model with the F-class, this was the class with the highest entropy. When the F-class had the highest entropy, it was assigned, on average, more than half the total weight (average weight = 0.59). Moreover, when fitting without the F-class, we noted that, in general, the class with the highest entropy is assigned a really large weight. For example, for all 5400 simulations (9 trees, 3 sequence lengths, 2 UDM distributions, and 100 repetitions per condition) when fitting the classes in either C20, C30, CK36, or C60, the class with the highest entropy had the largest weight, and on average, that weight was 4.71 times more than the weight assigned to that class when there is no misspecification. When fitting the classes in C40, the class with the second-highest entropy is the one with the largest weight, and it was also, on average, 3.99 times more than its weight when there

is no misspecification. Therefore we observe that this bias may not directly related to entropy but may instead be some other factor that is correlated with entropy.

We also explored the effects of misspecification of the exchangeabilities for data generated using the LG matrix but fitted the POISSON matrix. In this case, the tree estimation accuracy is also affected (for eg, OA values are in the range of 0.64 to 0.75 for sequence length of 1000 when it is correctly specified vs. a range of 0.60 to 0.71 when misspecified). Figure S3 in the Supplementary Material shows the average OA (A) and PM (B) for this case. In contrast to the reverse misspecification scenario (e.g. Figure 4), the F-class does not seem to hinder tree estimation. Figure S4 in the Supplementary Material, the analog of Figure 5, shows how, in this case, the weight of the F-class is close to zero, and therefore there is no likelihood difference with or without it. Furthermore, we did not find this phenomenon to be as strongly correlated to the Shannon entropy, as in the previous case. Although we found a shift in the correlation of entropy and class weight for all models. The average correlation between entropy and class weight for C60, C40, C30, and CK36 is 0.21 when there is no misspecification and -0.56 when there is. For C20 the shift is in a different direction, i.e the correlation between entropy and class weight is -0.12 when there is no misspecification and 0.11 when there is. Nonetheless, for this model and when there is misspecification, the class with the highest entropy is the one with the second lowest weight (average weight = 0.008). This same class has the highest weight when there is no misspecification (average weight = 0.142). This also suggests that entropy is somehow related to or affected by, model misspecification.

Finally, we note that it is formally possible that the generally good performance of CXX models in the foregoing analyses could be related to a tendency of these models to prefer topologies where long branches are apart (i.e. they could have a long-branch repulsion (LBR) bias). To test this, we simulated from a topology with long branches together, obtained from the tree in Figure 1 after swapping the clade composed of taxa $c1, ..., c_m, d1, ..., d_m$ together with the edge leading to it and the clade composed of taxa

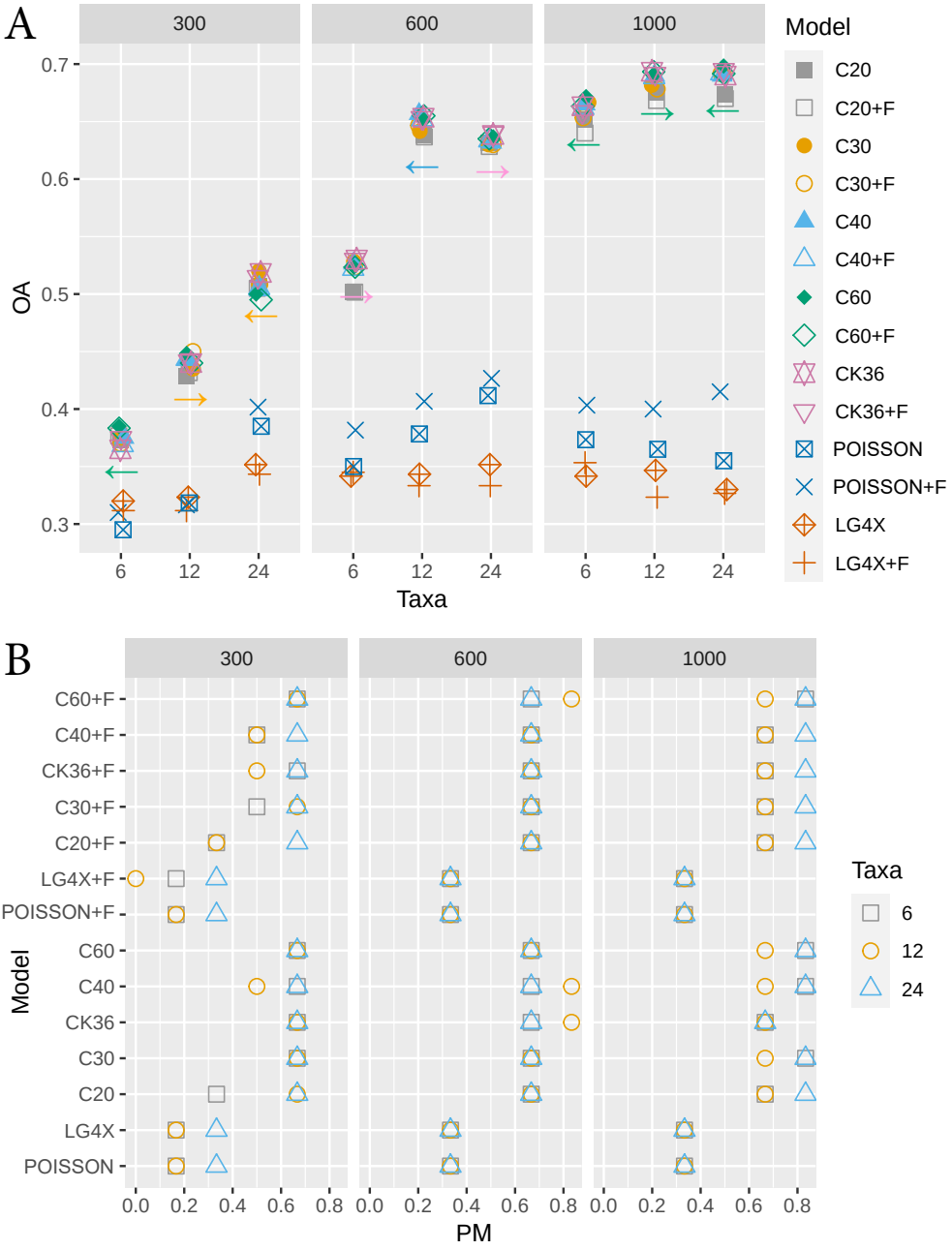

Fig. 3. (A) The plot of the OA values per model of the data generated under the UDM distributions and POISSON exchangeabilities. Different classes are fitted but in all cases, we fit POISSON exchangeabilities. The x-axis represents the number of taxa on the tree. The plot is divided by the sequence length. The lower bound for the 95% confidence interval (CI) of the model with the highest OA is depicted with an arrow whose color represents such model. An arrow pointing to the left represents CXX+F, an arrow pointing to the right represents CXX without F. (B) A similar plot to that on top but for PM. For this case, no confidence interval can be computed.

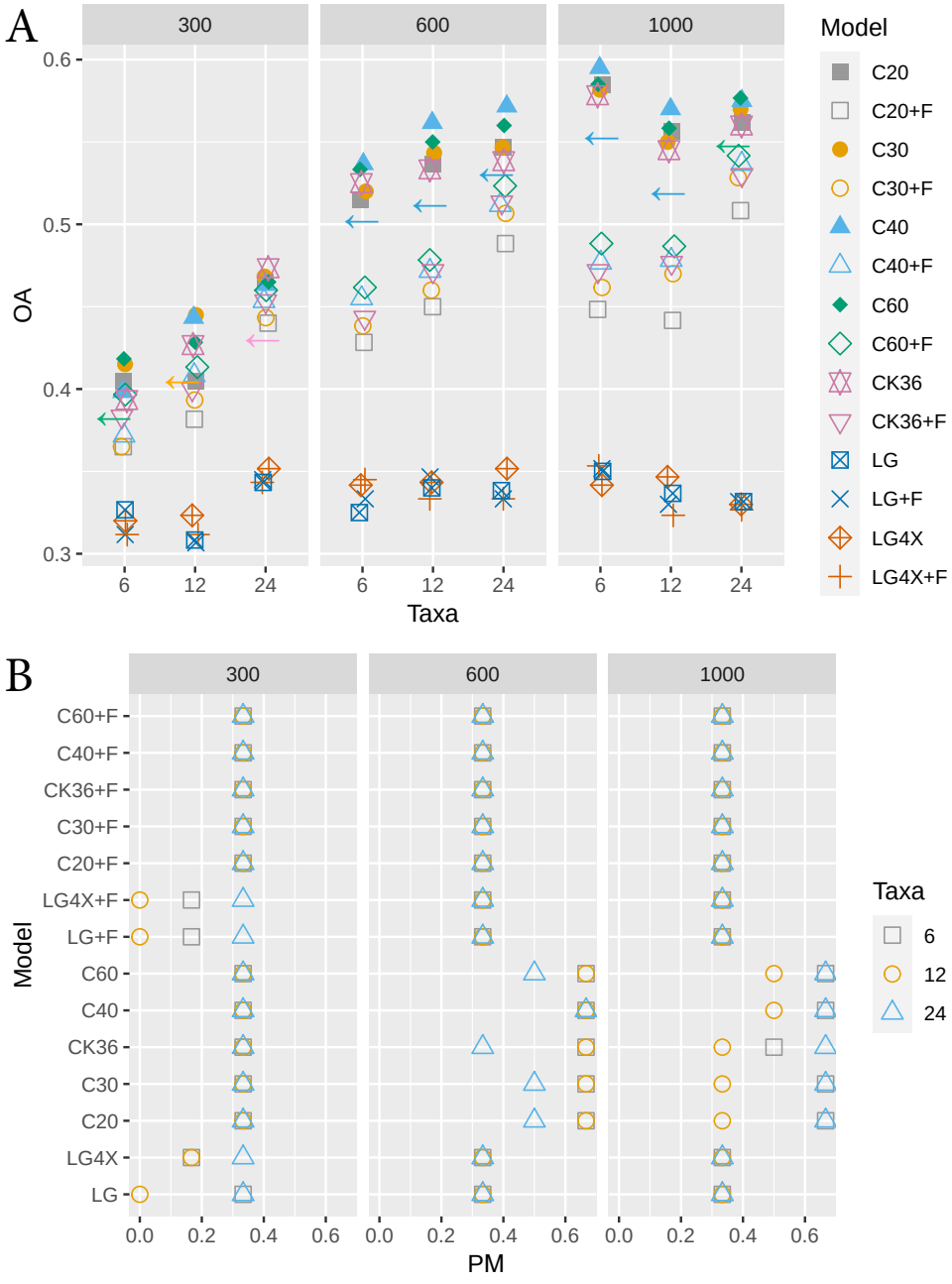

Fig. 4. (A) The plot of the OA values per model of data generated under the UDM distributions and POISSON exchangeabilities. Different classes are fitted but in all cases, we fit LG exchangeabilities. The x-axis represents the number of taxa on the tree. The plot is divided by the sequence length. The lower bound for the 95% confidence interval of the best model per scenario is depicted with a gray line. (B) A similar plot to that on top but for PM. For this case, no confidence interval can be computed.

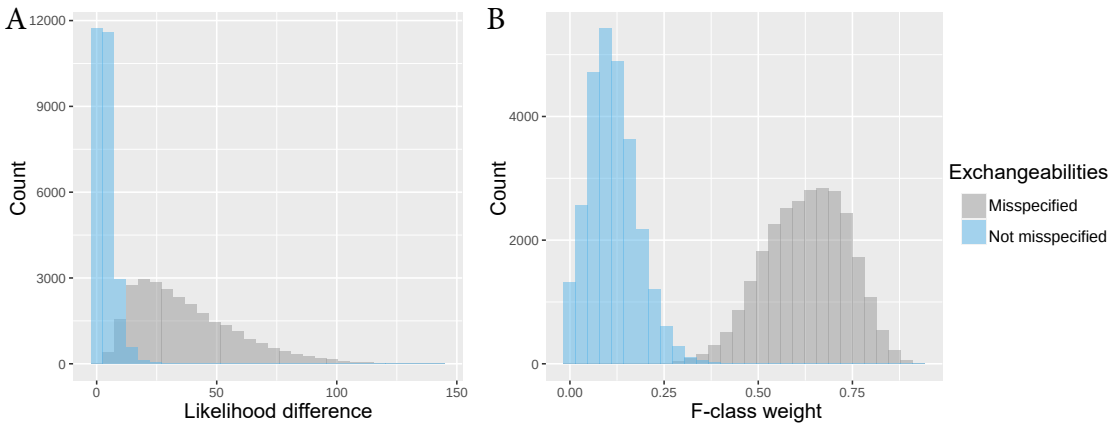

Fig. 5. (A) Histograms showing the difference between likelihoods values of models fitted with and without the F-class. No misspecification of the exchangeabilities is depicted in blue and in gray when there is. (B) Histograms showing the inferred F-class weight when there is no misspecification of the exchangeabilities (blue) and when there is (gray). The histograms consist of data generated under all trees, number of taxa, both UDM mixing distributions, and POISSON exchangeabilities. Misspecification of exchangeabilities refers to fitting using LG matrix instead of the POISSON matrix.

$e_1, ..., e_m$ together with the edge leading to it. Consequently, the long branches group together to the exclusion of short branches. Figure S5 in the Supplementary Material, gives the results for the 12 taxon case for this simulating scenario. On average, the estimates in OA and PM for the CXX models agree with those found when exploring LBA. We note that models with fewer classes tend to have better performance in these cases, suggesting these show a slight LBR bias. In contrast, the LG and LG4X models are strongly affected by LBR; i.e., they have notably poor performance under the LBA conditions and extremely good performance under the LBR conditions. A bias towards either the LBR or LBA topologies is not desirable in general. In this sense the CXX models, especially C30-C60, show little bias and are clearly better choices under these simulation conditions.

Real Data

To investigate the impact of mixture model choice on real data, we also analyzed three empirical data sets. These data sets are concatenated supermatrices:

I) a 133-protein dataset (24,291 sites $\times$ 40 taxa) assembled to assess the phylogenetic

position of the microsporidia in the tree of eukaryotes (Brinkmann et al. (2005)). The microsporidia are specifically related to Fungi but are sometimes recovered as branching outside of all eukaryotes because of an LBA artefact in which they are attracted to the outgroup archaeal sequences. We consider two trees: the correct tree recovered with the LG+C20+F+G model (T_I^{C20}) (Susko et al. (2018)) and the LBA tree recovered with the LG+F+G model (T_I^{LG})

II) a dataset of 146 proteins (35,371 sites $\times$ 37 taxa) assembled to assess the phylogenetic position of the nematodes in the animal tree of life (Lartillot et al. (2007)). In this case the two competing topologies are the correct topology (recovered with LG+C20+F+G: T_{II}^{C20}) where nematodes branch as sister to arthropods (i.e. the Ecdysozoa group) versus the artefactual topology recovered with LG+F+G (T_{II}^{LG}).

III) a dataset of 146 proteins (35,371 sites $\times$ 32 taxa) assembled to assess the phylogenetic position of the platyhelminths in the animal tree of life (Lartillot et al. (2007)). The correct position of platyhelminths within the Protostomia is reflected in the tree recovered by CAT+GTR (T_{III}^{CAT}) instead of the artefactual Coelomata topology (T_{III}^{LG}) recovered by LG+F+G and many mixture models (see Lartillot et al. (2007), Susko et al. (2018) and Wang et al. (2017))

For each of these trees, we computed the log-likelihoods when fitting classes in C20, C40, and C60, with and without the F-class, and with both LG and POISSON matrices. These likelihoods are shown in Table 4. In all cases, fitting with the LG matrix produces higher likelihood values.

For data set I, we observe that the correct tree is obtained with the largest log-likelihood differences over the incorrect tree when C60 and C20 are fit with LG exchangeabilities. For this data set the F-class only sometimes negatively affects topological estimation, but never improves it. For data set II, POISSON exchangeabilities strongly favor the correct topology over the incorrect one relative to LG. Here the F-class makes little difference but again never increases support for the correct tree. For data set

Model	**Fitted**	$L(T_I^{C20})$	$D(T_I^{C20})$	$L(T_{II}^{C20})$	$D(T_{II}^{C20})$	$L(T_{III}^{CAT})$	$D(T_{III}^{CAT})$
C60	**LG+F**	-715744	14	-712614	24	-626777	-2
C60	**LG**	-716579	20	-713137	27	-627241	-1
C60	**POI+F**	-722917	13	-718761	60	-631550	29
C60	**POI**	-722917	13	-718761	60	-631550	29
C40	**LG+F**	-716584	4	-713345	26	-627478	-6
C40	**LG**	-717555	9	-713882	30	-627977	-5
C40	**POI+F**	-724391	9	-720761	58	-633321	26
C40	**POI**	-724391	9	-720761	58	-633321	26
C20	**LG+F**	-718315	7	-714772	19	-628597	-10
C20	**LG**	-719775	19	-715659	23	-629393	-8
C20	**POI+F**	-727735	9	-723988	56	-635979	17
C20	**POI**	-727737	9	-723988	56	-635979	17

Table 4. The log-likelihoods of the trees estimated from the empirical data sets, where $D(T_J)$ denotes the log-likelihood of the 'correct tree' (e.g. C20 or CAT superscripts) minus the 'incorrect' tree (e.g. LG superscripts). POI stands for POISSON matrix of exchangeabilities.

III, POISSON exchangeabilities favor the correct tree over the incorrect tree, with C60 showing the biggest log-likelihood difference. LG exchangeabilities seem to always favor the incorrect tree.

Overall, the F-class never increases support for the correct tree and sometimes decreases it. Whether LG improves estimation versus POISSON depends on the data set. However, we note that the proteins and taxa in datasets II and III heavily overlap so the outcomes of these analyses are not technically independent.

Discussion

By extending earlier results (Kiefer and Wolfowitz (1956); Lindsay (1983); Yourdkhani et al. (2021)), we confirm that, for profile mixture models, the tree and mixing parameters of profile mixture models are statistically consistent even with a large number of classes. However, since good performance is not guaranteed for short alignments, we conducted an extensive simulation study of the performance and properties of profile mixture models with smaller data sets with the goal of determining if over-parameterization was a problem. We also investigated the effects of model

misspecification through both misspecification of the frequency classes and the exchangeabilities. Finally, the effects of the F-class was also investigated in all possible settings. These analyses provide useful theoretical and practical insights regarding model fit. Our main findings are the following:

(A) ***Over-parameterization is not a problem for complex models:*** For all alignment sizes explored here, we saw no evidence (in terms of MISE, MD, OA and PM) that use of more complex profile mixture models led to more variable or poorer estimation. This is true even for models having several classes with zero weights estimates. Consistent with the theoretical results for large numbers of sites, over-parameterization of these mixture models does not appear to be a problem for shorter alignments.

Since it is the mixture structure that is important in assessing whether models are overparameterized, large sample results likely extend to rates-across-sites mixtures (Yang (1994); Felsenstein and Churchill (1996); Mayrose et al. (2005); Susko et al. (2003)) and the types of mixtures used to infer selection pressure (Yang et al. (2000)). We also speculate that some of the small sample results found here may extend to those settings too but additional work is needed.

(B) ***Misspecification of the frequency vectors does not necessarily imply bad fit:*** Misspecification of the frequency vectors in profile mixture models does not cause problems if the estimated CDF can adequately approximate the true CDF. The more data available, the more classes are likely needed to closely approximate the true CDF.

(C) ***Simple models behave poorly:*** Likely as a consequence of (B), both the site-homogeneous POISSON and LG models, and the site-heterogeneous LG4X model perform very poorly in all scenarios. We believe this is because these have one (LG and POISSON) or very few (LG4X) classes. Although inference using simple models

607 can be much faster we do not recommend their use given their poor performance (i.e.
608 susceptibility to LBA) under realistic site-heterogeneous simulation conditions.

609 (D) ***Misspecification of exchangeabilities and the presence of an F-class can***
610 ***severely affect tree estimation:***

611 A severe decrease in accuracy of tree estimation is observed for data generated under
612 the POISSON matrix and fitted using the LG matrix. In this scenario, it is clear that
613 the F-class degrades performance. Such misspecification resulted in large weights of
614 the F-class. For data generated under the LG matrix and fitted using the POISSON
615 matrix, performance is affected but not as severely as in the previous case.

616 From this, we hypothesize that misspecification of exchangeabilities is more
617 problematic when the matrix used to fit is less uniform than the true exchangeability
618 matrix. In that case, the F-class tends to be accorded a large weight that leads to
619 poorer tree estimation performance. Although the reverse misspecification scenario
620 also degrades performance somewhat, the F-class has a little role in that case. We
621 suspect that because the LG matrix was originally estimated as an "approximation"
622 of a GTR matrix for many alignments in a site-homogeneous context, the LG matrix
623 is less uniform than it would be if it was estimated in the presence of profile mixture
624 models like the CXX set. Thus, we suspect the pathological behavior of the F-class
625 and poor performance may apply to real estimation settings. We note that use of
626 both the LG matrix and the F-class lead to higher likelihoods, so model selection
627 criteria like AIC will frequently favor their use in real settings. Since the F-class
628 never appeared to improve estimation in any of the simulations or real data analysis
629 settings we examined, we discourage its use in site-profile mixture models.

630 (E) ***Better likelihood estimates do not imply better tree estimates:*** As a
631 consequence of (D), and also observed in the data, better estimates in likelihood do
632 not imply better tree estimates. This is also weakly observed even when there is no

misspecification of the exchangeabilities.

(F) ***Adding more taxa can improve or hurt tree estimation accuracy:*** Adding more taxa generally improves MISE and tree estimation. Surprisingly, when the model is misspecified (e.g. using UDM frequencies and misspecification of exchangeabilities) adding taxa does not always improve estimation; in one case it actually decreases performance (Fig. 4).

The poor performance of the methods when exchangeabilities are misspecified provides a strong motivation to develop software tools that allow ML estimation of a GTR matrix over all sites in the presence of a profile mixture model. In future work, we plan to construct mixing distributions that closely approximate the true CDF for data, hoping this would lead to more accurate tree estimation than current models.

To finalize, we give some practical recommendations for single gene phylogeny inference. First, we do not discourage the use of 'rich models' (those with many frequency classes), even when several classes have zero weight estimates. We suggest avoiding models with one or very few frequency classes. We also discourage the use of the F-class, unless both scenarios, with and without the F-class, can be explored.

Funding

This work and H.B. were supported by the Moore-Simons Project on the Origin of the Eukaryotic Cell, Simons Foundation grant 735923LPI (DOI: `https://doi.org/10.46714/735923LPI`) and by NSERC Discovery Grants awarded to A.J.R. and E.S.

References

Anderson, F. E. and A. R. Lindgren. 2021. Phylogenomic analyses recover a clade of large-bodied decapodiform cephalopods. Molecular Phylogenetics and Evolution

156:107038.

Billera, L. J., S. P. Holmes, and K. Vogtmann. 2001. Geometry of the space of phylogenetic trees. Advances in Applied Mathematics 27:733–767.

Brinkmann, H., M. van der Giezen, Y. Zhou, G. P. de Raucourt, and H. Philippe. 2005. An Empirical Assessment of Long-Branch Attraction Artefacts in Deep Eukaryotic Phylogenomics. Systematic Biology 54:743–757.

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

Daubin, V. 2002. A phylogenomic approach to bacterial phylogeny: Evidence of a core of genes sharing a common history. Genome Research 12:1080–1090.

Dufayard, J.-F., L. Duret, S. Penel, M. Gouy, F. Rechenmann, and G. Perrière. 2005. Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. Bioinformatics 21:2596–2603.

Felsenstein, J. 1978. Cases in which Parsimony or Compatibility Methods will be Positively Misleading. Systematic Biology 27:401–410.

Felsenstein, J. and G. A. Churchill. 1996. A Hidden Markov Model approach to variation among sites in rate of evolution. Molecular Biology and Evolution 13:93–104.

Feuda, R., M. Dohrmann, W. Pett, N. Lartillot, G. Wörheide, and D. Pisani. 2017. Improved Modeling of Compositional Heterogeneity Supports Sponges as Sister to All Other Animals. Current Biology 27:3864–3870.

Franzosa, E. A. and Y. Xia. 2009. Structural Determinants of Protein Evolution Are Context-Sensitive at the Residue Level. Molecular Biology and Evolution 26:2387–2395.

Gaston, D., E. Susko, and A. J. Roger. 2011. A phylogenetic mixture model for the identification of functionally divergent protein residues. Bioinformatics 27:2655–2663.

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

684 Goldstein, R. A. 2008. The structure of protein evolution and the evolution of protein
685 structure. Current Opinion in Structural Biology 18:170–177.

686 Halpern, A. L. and W. J. Bruno. 1998. Evolutionary distances for protein-coding sequences:
687 modeling site-specific residue frequencies. Molecular Biology and Evolution 15:910–917.

688 Hartigan, J. A. and M. A. Wong. 1979. Algorithm as 136: A k-means clustering algorithm.
689 Journal of the Royal Statistical Society. Series C (Applied Statistics) 28:100–108.

690 Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation
691 data matrices from protein sequences. Bioinformatics 8:275–282.

692 Kiefer, J. and J. Wolfowitz. 1956. Consistency of the maximum likelihood estimator in the
693 presence of infinitely many incidental parameters. The Annals of Mathematical
694 Statistics 27:887–906.

695 Lanfear, R., P. B. Frandsen, A. M. Wright, T. Senfeld, and B. Calcott. 2016.
696 PartitionFinder 2: New Methods for Selecting Partitioned Models of Evolution for
697 Molecular and Morphological Phylogenetic Analyses. Molecular Biology and Evolution
698 34:772–773.

699 Lartillot, N., H. Brinkmann, and H. Philippe. 2007. Suppression of long-branch attraction
700 artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol
701 7:S4.

702 Lartillot, N. and H. Philippe. 2004. A Bayesian Mixture Model for Across-Site
703 Heterogeneities in the Amino-Acid Replacement Process. Molecular Biology and
704 Evolution 21:1095–1109.

705 Lartillot, N., N. Rodrigue, D. Stubbs, and J. Richer. 2013. PhyloBayes MPI: Phylogenetic

706 Reconstruction with Infinite Mixtures of Profiles in a Parallel Environment. Systematic
707 Biology 62:611–615.

708 Le, S. Q., C. C. Dang, and O. Gascuel. 2012. Modeling Protein Evolution with Several
709 Amino Acid Replacement Matrices Depending on Site Rates. Molecular Biology and
710 Evolution 29:2921–2936.

711 Le, S. Q. and O. Gascuel. 2008a. An Improved General Amino Acid Replacement Matrix.
712 Molecular Biology and Evolution 25:1307–1320.

713 Le, S. Q. and O. Gascuel. 2008b. An Improved General Amino Acid Replacement Matrix.
714 Molecular Biology and Evolution 25:1307–1320.

715 Le, S. Q. and O. Gascuel. 2010. Accounting for Solvent Accessibility and Secondary
716 Structure in Protein Phylogenetics Is Clearly Beneficial. Systematic Biology 59:277–287.

717 Li, Y., X.-X. Shen, B. Evans, C. W. Dunn, and A. Rokas. 2021. Rooting the Animal Tree
718 of Life. Molecular Biology and Evolution Msab170.

719 Lindsay, B. G. 1983. The Geometry of Mixture Likelihoods: A General Theory. The
720 Annals of Statistics 11:86 – 94.

721 Ly-Trong, N., S. Naser-Khdour, R. Lanfear, and B. Q. Minh. 2021. Alisim: A fast and
722 versatile phylogenetic sequence simulator for the genomic era. bioRxiv .

723 Massey, F. J. 1951. The Kolmogorov-Smirnov test for goodness of fit. Journal of the
724 American Statistical Association 46:68–78.

725 Mayrose, I., N. Friedman, and T. Pupko. 2005. A Gamma mixture model better accounts
726 for among site rate heterogeneity. Bioinformatics 21:ii151–ii158.

727 Minh, B. Q., H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von
728 Haeseler, and R. Lanfear. 2020. IQ-TREE 2: New Models and Efficient Methods for
729 Phylogenetic Inference in the Genomic Era. Molecular Biology and Evolution
730 37:1530–1534.

731 Pál, C., B. Papp, and M. J. Lercher. 2006. An integrated view of protein evolution. Nature
732 Reviews Genetics 7:337–348.

733 Philippe, H., H. Brinkmann, D. V. Lavrov, D. T. J. Littlewood, M. Manuel, G. Wörheide,
734 and D. Baurain. 2011. Resolving difficult phylogenetic questions: Why more sequences
735 are not enough. PLOS Biology 9:1–10.

736 Philippe, H. and J. Laurent. 1998. How good are deep phylogenetic trees? Current Opinion
737 in Genetics and Development 8:616–623.

738 Pisani, D., W. Pett, M. Dohrmann, R. Feuda, O. Rota-Stabelli, H. Philippe, N. Lartillot,
739 and G. Wörheide. 2015. Genomic data do not support comb jellies as the sister group to
740 all other animals. Proceedings of the National Academy of Sciences 112:15402–15407.

741 Pupko, T., D. Huchon, Y. Cao, N. Okada, and M. Hasegawa. 2002. Combining Multiple
742 Data Sets in a Likelihood Analysis: Which Models are the Best? Molecular Biology and
743 Evolution 19:2294–2307.

744 Raymann, K., C. Brochier-Armanet, and S. Gribaldo. 2015. The two-domain tree of life is
745 linked to a new root for the archaea. Proceedings of the National Academy of Sciences
746 112:6670–6675.

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

749 Schrempf, D., N. Lartillot, and G. Szöllősi. 2020. Scalable Empirical Mixture Models That
750 Account for Across-Site Compositional Heterogeneity. Molecular Biology and Evolution
751 37:3616–3631.

752 Scott, D. W. 1992. Multivariate density estimation theory, practice, and visualization. J.
753 Wiley.

754 Self, S. G. and K.-Y. Liang. 1987. Asymptotic properties of maximum likelihood

estimators and likelihood ratio tests under nonstandard conditions. Journal of the American Statistical Association 82:605–610.

Simion, P., H. Philippe, D. Baurain, N. King, G. Wörheide, and M. Manuel. 2017. A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals. Current Biology 27:958–967.

Susko, E., C. Field, C. Blouin, and A. J. Roger. 2003. Estimation of Rates-Across-Sites Distributions in Phylogenetic Substitution Models. Systematic Biology 52:594–603.

Susko, E., L. Lincker, and A. J. Roger. 2018. Accelerated Estimation of Frequency Classes in Site-Heterogeneous Profile Mixture Models. Molecular Biology and Evolution 35:1266–1283.

Thorndike, R. L. 1953. Who belongs in the family? Psychometrika 18:267–276.

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

Wang, H.-C., B. Q. Minh, E. Susko, and A. J. Roger. 2017. Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. Systematic Biology 67:216–235.

Wang, H.-C., E. Susko, and A. J. Roger. 2014. An Amino Acid Substitution-Selection Model Adjusts Residue Fitness to Improve Phylogenetic Estimation. Molecular Biology and Evolution 31:779–792.

Whelan, S. and N. Goldman. 2001. A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. Molecular Biology and Evolution 18:691–699.

Wickett, N. J., S. Mirarab, N. Nguyen, T. Warnow, E. Carpenter, N. Matasci, S. Ayyampalayam, M. S. Barker, J. G. Burleigh, M. A. Gitzendanner, B. R. Ruhfel,

E. Wafula, J. P. Der, S. W. Graham, S. Mathews, M. Melkonian, D. E. Soltis, P. S. Soltis, N. W. Miles, C. J. Rothfels, L. Pokorny, A. J. Shaw, L. DeGironimo, D. W. Stevenson, B. Surek, J. C. Villarreal, B. Roure, H. Philippe, C. W. dePamphilis, T. Chen, M. K. Deyholos, R. S. Baucom, T. M. Kutchan, M. M. Augustin, J. Wang, Y. Zhang, Z. Tian, Z. Yan, X. Wu, X. Sun, G. K.-S. Wong, and J. Leebens-Mack. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. Proceedings of the National Academy of Sciences 111:E4859–E4868.

Williams, T. A., P. G. Foster, C. J. Cox, and T. M. Embley. 2013. An archaeal origin of eukaryotes supports only two primary domains of life. Nature 504:231–236.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. Journal of Molecular Evolution 39:306 – 314.

Yang, Z. 1996. Maximum-Likelihood Models for Combined Analyses of Multiple Sequence Data. J Mol Evol 42:587–596.

Yang, Z., R. Nielsen, N. Goldman, and A.-M. K. Pedersen. 2000. Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites. Genetics 155:431–449.

Yourdkhani, S., E. S. Allman, and J. A. Rhodes. 2021. Parameter Identifiability for a Profile Mixture Model of Protein Evolution. Journal of Computational Biology 28:570–586.

Appendix

Statistical consistency of the MLE

As mentioned in the section entitled *"Mixture models and over-parameterization with large samples"*, the results of Kiefer and Wolfowitz (1956) imply that the tree and

mixing parameters are consistent. In this section we elaborate on why this is true in the context of site-profile mixture models. To do this, we first introduce the concept of a *forest*. This allows us to then extend the parameter space of the model (to a compact one) so the regularity conditions in Kiefer and Wolfowitz (1956) are satisfied.

A phylogenetic *forest* F on X is a collection of phylogenetic trees F_q on X_q, known as *components*, where $X = \cup X_q$ and $X_q \cap X_p = \emptyset$ for any two components.

We extend the parameter space Θ defined in Section *"Mixture models and over-parameterization with large samples"* by substituting point (i) of that section with:

(i') A metric forest F on N taxa obtained after removing a, possibly empty, set of edges from a rooted metric binary tree on N taxa and retaining only components that display taxa.

In this case, the substitution process of a profile mixture model is as follows: for each site, a class $\boldsymbol{\pi}_c$ is sampled with probability w_c, and a rate parameter r_k is sampled with probability d_k. On each component F_q of F an independent substitution process on F_q with exchangeabilities R, root distribution $\boldsymbol{\pi}_c$, and a rate parameter r_k is conducted. For a given site pattern $\boldsymbol{x}_i$, the likelihood function is determined by a weighted average of the product of partial site likelihoods conditional on each site-profile class and site-rate class per component:

$$L(\theta|\boldsymbol{x}_i) = \sum_{c=1}^{C} w_c \sum_{k=1}^{K} d_k \prod_{q=1}^{Q} P(\boldsymbol{x}_i|F_q, R, \boldsymbol{\pi}_c, r_k),$$

where $\theta = (F, R, \{\boldsymbol{\pi}_c\}, \{w_c\}, \{r_k\}, \{d_k\}) \in \Theta$ and Q are the number of components of F. The reasoning behind forests is to account for infinite edge lengths, which are represented by the edges missing from the tree defining F. This not only allows us to consider this limiting case but also, it weakly depicts the effects of functional divergence (Gaston et al. (2011)). Note that when no edges are removed from the tree in (i'), the resulting forest has one component and the likelihood is the same as the one defined in the section *"Mixture models and over-parameterization with large samples."*

We now show we "compactified" the parameter space Θ of the profile mixture. We show this by proving (i') above, and elements (iii) and (iv) of parameter space Θ are compact. We do not consider (ii) (i.e. the matrix of exchangeabilities) in the parameter space because, for the models we consider here, it is fixed beforehand.

Clearly (iii) is compact since both, the root distribution vectors and the class weights are closed and bounded. To argue (i') is compact, we need to recall that the space of rooted metric phylogenetic trees on n taxa can be viewed as a collection of $(2n-3)!!$ open cubes corresponding to all different tree topologies (Billera et al. (2001)). The limiting cases in these cubes correspond to infinite edge lengths on the trees. We can ensure boundedness by re-parameterizing edge lengths via the logistic function $p = \frac{e^t}{1+e^t}$. Cases with an edge length $p = 1$ corresponds to those edge lengths being infinite. The limiting likelihoods in those cases correspond to forests. Therefore (i') can be viewed as a compact space.

For (iv), the rate weights are clearly compact (closed and bounded). Now, even if the rate parameters are unbounded from above, the limiting case, i.e. when $r \to \infty$, is equivalent to the process occurring in the forest where all components are just single taxa. Therefore Θ is compact, and the results of Kiefer and Wolfowitz (1956) hold in our context.

Another important statement mentioned in the section *"Mixture models and over-parameterization with large samples"* is: even when the parameter estimation is unrestricted and any mixing distribution is allowed, the maximum likelihood estimator will be a finite mixing distribution. This is an implication from Theorem 3.1 in Lindsay (1983). For this result to hold, the trace of the likelihood curve over the mixing parameters must be compact. This follows immediately from the fact that: (1) the mixing parameter space is compact; and (2) the image of continuous functions, such as the trace of the likelihood curve, is compact whenever the domain is compact.

Identifiability

In this section we argue why, as mentioned in the section *"Mixture models and over-parameterization with large samples,"* for many of the cases considered here the tree parameter is identifiable.

In Theorem 5.7 in Yourdkhani et al. (2021) it is shown that for profile mixture models with $C \cdot K < 72$, where C is the number of classes and K the number of rates, and more than 8 taxa, the tree and numerical parameters are generically identifiable, up to arbitrary re-scaling of the tree and the exchangeability matrix. Generically identifiable means identifiable except maybe in a set of measure zero; informally speaking this means identifiable except maybe in a tiny subset of parameters relative to the full parameter space. Although in such work there is no description of the generic setting of the parameter space, we argue that with just a small perturbation of the parameters one can always guarantee the result to hold.

Since in all models considered here fixed parameters are obtained empirically (that is these have no structure for eg. be solutions of a phylogenetic invariant) and the parametric function is continuous, there exists $\boldsymbol{\epsilon}$ such that a translation of the numerical parameters by $\boldsymbol{\epsilon}$ will make these generic. This is true for the frequency vectors, weights, rate parameters, edge lengths, and the exchangeabilities matrices. While the POISSON matrix is not generated from data, the proof of Theorem 5.7 in Yourdkhani et al. (2021) is built on this matrix up to a constant and therefore identifiability also holds in this case.