Accelerated Estimation of Frequency Classes in Site-heterogeneous Profile Mixture Models

Edward Susko[1] Léa Lincker[2,3] and Andrew J. Roger[3]

[1] *Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia*

[2] *École nationale supérieure de Techniques Avancées, Palaiseau, France*

[3] *Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia, Canada B3H 4H7*

Corresponding Author: Edward Susko, Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, Canada B3H 3J5; Phone: (902) 494-8865; Fax: (902) 494-5130; E-mail: susko@mathstat.dal.ca

Abstract

As a consequence of structural and functional constraints, proteins tend to have site-specific preferences for particular amino acids. Failing to adjust for heterogeneity of frequencies over sites can lead to artefacts in phylogenetic estimation. Site-heterogeneous mixture-models have been developed to address this problem. However, due to prohibitive computational times, maximum likelihood implementations utilize fixed component frequency vectors inferred from sequences in a database that are external to the alignment under analysis. Here we propose a composite likelihood approach to estimation of component frequencies for a mixture model that directly uses the data from the alignment of interest. In the common case that the number of taxa under study is not large, several adjustments to the default composite likelihood are shown to be necessary. In simulations, the approach is shown to provide large improvements over hierarchical clustering. For empirical data, substantial improvements in likelihoods are found over mixtures using fixed components.

Key words: mixture model, site-specific model, phylogenetics, protein models

Introduction

Phylogenomic methods that analyze large numbers of orthologous genes are increasingly used to resolve deep phylogenetic divergences in the tree of life (Brown et al. 2013; Wickett et al. 2014; Pisani et al. 2015). Variability of estimation decreases with larger alignment lengths but computational cost increases substantially and the possibility of systematic bias makes it important to accurately model the underlying amino acid substitution process (Philippe et al. 2011).

Amino acid substitutions are usually modeled as occurring independently at sites in an alignment and according to a Markov process along a tree. The most common approach assumes a constant rate matrix throughout the tree, determined by stationary frequencies that are constant across sites and an empirically derived exchangeability matrix. Exchangeabilities are fixed in advance of analyses and empirically determined. Examples include the JTT exchangeability matrix (Jones, Taylor and Thornton 1992), the WAG matrix (Whelan and Goldman 2001) and the LG matrix of Le and Gascuel (2008). Frequencies are usually determined from the alignment as the observed frequencies of amino acids over all sites and taxa. Allowance is made for heterogeneity of substitution processes over sites through a mixture model of rates for sites, arising from a discretized Gamma distribution (Yang 1994).

Rate heterogeneity, however, is not usually sufficient to adjust for the differing structural or functional constraints that lead to different sites having different preferences for specific amino acids. Many sites appear to allow a relatively limited set of amino acids with the nature of the set often varying across sites (Halpern and Bruno 1998; Lartillot and Phillipe 2004). For instance, some sites show a restricted alphabet of amino acids. This restricted alphabet of amino acids is determined by the structural and functional constraints at those sites in the protein. Such constraints can include requirements for amino acids that are hydrophobic and aliphatic (e.g. V, I, and L), aromatic (e.g. F, Y and W), acidic (e.g. D and E), basic (e.g. R, K and H) or with other biophysical properties. Generally, frequencies at sites are often less uniform than those predicted by the observed aggregate frequencies over sites that are used in conventional models (Lartillot et al. 2007; Wang et al. 2008). Sites with highly skewed composition tend to become saturated with changes at smaller evolutionary distances than when frequencies are uniform. At such sites, patterns for subsets of taxa that are not very distantly related can appear similar to those of subsets of taxa that are distantly related, leading to an underestimation of relative evolutionary divergences (large to small). As a consequence, conventional models have shown a tendency towards long-branch attraction (LBA) biases in phylogenetic estimation (Lartillot et al. 2007; Wang et al. 2008). By contrast, site-heterogeneous models effectively downweight the importance of sites with highly skewed composition to the inference of relative divergences, implicitly recognizing that saturation is a potential explanation for the patterns at these sites.

To adjust for heterogeneity of frequencies over sites, mixture models (Lartillot and Philippe 2004; Wang et al. 2008; Le et al. 2008) and partitioned models (Yang 1996; Pupko et al. 2002; Lanfear et al. 2012) have been developed. We focus attention on the mixture approach here. A widely used class of mixture models implemented according to Bayesian principles are the CAT models of Lartillot et al. (2013). The mixing distribution of stationary frequencies over sites is modeled without restriction using a Dirichlet process model. While CAT models have been shown to fit data better and alleviate LBA bias (Lartillot et al. 2007), a concern has been that the Markov chain Monte Carlo computational techniques used to approximate posterior probabilities can suffer from convergence difficulties with large data (Whelan et al. 2015; Pisani et al. 2015; Whelan and Halanych 2016).

Due to their substantial additional computational burden, maximum likelihood (ML) models of mixtures of frequencies (Wang et al. 2008; Le et al. 2012) assume a mixing distribution with component frequency vectors that are fixed in advance, rather than being estimated from the data. Several sets of fixed frequency vectors are available (Wang et al. 2008, 2014; Le et al. 2008), each having been determined from differing previous empirical data and using different methods. Simulations and empirical studies have shown that these mixture models are more robust to LBA bias than a model that uses a single stationary frequency vector (Wang et al. 2017).

When applied to new data, current mixture models (Wang et al. 2008; Le et al. 2012) utilize fixed frequency profiles determined from external data. Consequently, these fixed frequency profiles may not be consistent with the actual frequency profiles present in the data under consideration. In theory, frequency profiles can be included among the parameters optimized in ML estimation. Because each profile has 20 elements such an approach increases the dimension of the optimization problem substantially. Moreover, whereas the rate matrices using fixed frequency profiles remain fixed throughout optimization, each new frequency profile considered when they are being optimized requires additional eigenvalue decompositions in order to calculate substitution matrices from the rate matrices through matrix exponentiation. Finally, by contrast with derivatives for edge-lengths which can be calculated efficiently using single sweeps of the pruning algorithm of Felsenstein (1981), repeated pruning algorithm applications are required to approximate derivatives for each element of each frequency profile. The increase in computation complexity of likelihood evaluations and the difficulties with derivative calculations, renders the approach prohibitive in practice.

In this study we investigate feasible methods for estimating component frequency vectors in the mixing distribution. Through simulations the main methods are shown to provide substantial improvements over hierarchical clustering. For empirical data, substantial improvements in likelihoods are found over mixtures using fixed pre-determined components.

Theory and Methods

Conventional and Frequency Mixture Models

Throughout this article we assume a mixture-of-frequencies model generated the alignment. Let $x_1, \ldots, x_n$ denote the columns of the alignment. Here x_i is a site pattern, $x_{i1}, \ldots, x_{im}$, where x_{is} is the amino acid at site i for taxa s, $s = 1, \ldots, m$. For instance $x_i = AAAR$ denotes that the first three taxa were observed to have amino acid A and the fourth taxon, amino acid R. As with conventional models we assume the x_i are independent. Consequently the log likelihood for the data is of the form

$$l(\theta) = \sum_{i=1}^{n} \log p(x_i; \theta),$$

where θ denotes all unknown parameters in the model and $p(x_i; \theta)$ the probability of observing site pattern x_i at a site.

A conventional site-homogeneous model provides the basis for the frequency profile mixture model. In the site-homogeneous model, at a site, evolution along any edge in the tree occurs according to a Markov chain with rate matrix Q. In conventional models the rate matrix is parameterized as $Q_{ij} = S_{ij}\pi_j$, for $i \neq j$, where the S_{ij} are fixed exchangeability parameters determined from emprical data. Common choices include the JTT matrix (Jones, Taylor and Thornton 1992), the WAG matrix (Whelan and Goldman 2001) or the LG matrix of Le and Gascuel (2008). We use the LG matrix throughout most of this article. With Q_{ij} parameterized as $Q_{ij} = S_{ij}\pi_j$, the π_j are interpretable as the stationary frequencies of amino acids under the model. For conventional models, usually π_j is estimated by the observed aggregate frequency of amino acid j over all taxa and sites. In such cases the model nomenclature is JTT+F or LG+F depending on the exchangeability matrix.

Rate variation is allowed in a conventional model through a finite gamma mixture model described in Yang (1994). In that model rates arise independently from a discrete distribution that assigns probability $1/K$ to rates $r_1, \ldots, r_K$; in all applications we use $K = 4$. The r_k are chosen to provide a discrete approximation to a gamma distribution and depend on the shape parameter α for that gamma distribution. For a site having rate r_k, the probability of observed data x can be directly calculated using the pruning algorithm of Felsentsein (1981). We denote that probability as $p(x|k; \zeta, \pi)/K$, where we have indicated dependence on the stationary frequencies π and the other parameters ζ, which includes α, the unrooted topology τ and edge lengths $\boldsymbol{t}$. Since the r_k are unobserved, the unconditional probability of the data actually observed is $p(x; \zeta, \pi) = \sum_k p(x|k; \zeta, \pi)/K$. The model nomenclature for discrete gamma rate variation, is JTT+F+Γ or LG+F+Γ depending on the exchangeability matrix.

The frequency mixture model is a mixture on top of the gamma mixture. Frequency vectors arise independently from a discrete distribution that assigns probabilities $w_1, \ldots, w_C$

to frequency vectors $\pi^{(1)}, \ldots, \pi^{(C)}$. Similarly as for the gamma mixture, the probability of observing x at a site is $p(x;\theta) = \sum_c w_c p(x;\zeta,\pi^{(c)})$, where now $p(x;\zeta,\pi^{(c)})$ denotes the conditional probability of x given the frequency vector $\pi^{(c)}$ for the site. The parameter vector θ includes the parameters ζ that are present in a conventional model and also the frequency vectors and weights, w_c. In current applications of frequency mixture models, the w_c are estimated from the data but the frequency vectors, $\pi^{(1)}, \ldots, \pi^{(C)}$, are fixed and do not necessarily reflect the amino acid preferences at sites for the data at hand. A particular choice that will be utilized in what follows are the C-series frequency vectors from Le et al. (2008) which are a set of C frequency vectors derived from empirical data, $C = 10, 20, \ldots, 60$. The model is frequently applied with a discrete gamma rate model and an additional frequency vector, determined as the observed frequencies of the amino acids, and is denoted as LG+C20+F+Γ for an LG exchangeability matrix and when $C = 20$.

Composite Likelihood - The Multinomial Mixture Likelihood

By comparison with conventional models, the frequency mixture model includes additional weight parameters, $w_1, \ldots, w_C$ and frequency vectors $\pi^{(1)}, \ldots, \pi^{(C)}$. Estimation of the weights through ML requires a relatively minor additional computational cost. Estimation of frequency vectors, however, is usually prohibitive. First, because each frequency vector introduces 19 additional parameters, there are $19C$ additional parameters requiring optimization; with $C = 20$ components, for instance, this gives 380 additional parameters. Second, the new frequency vectors encountered in the course of optimization give new rate matrices which result in completely different $p_{ij}(t)$ along all edges of the tree. As a consequence, likelihood evaluation for a new set of frequency vectors requires a completely new application of the pruning algorithm. By contrast, edge-length estimation, which similarly involves relatively large numbers of parameters, can re-use previous calculations to more efficiently calculate likelihoods. Finally, optimization requires derivative calculations. Whereas derivatives of log likelihoods for edge-lengths can be calculated exactly and efficiently, derivatives for frequency parameters need to be approximated.

Even in the case of a conventional model where there is only a single frequency vector, estimation is not usually through ML. Instead the observed aggregate frequencies of amino acids over sites and taxa are directly calculated. An alternative characterization of the observed frequencies is that they are ML estimates of the frequencies for a tree with infinite edge lengths. The log likelihood in this case is a multinomial log likelihood,

$$\sum_i \log(\prod_s \pi_{x_{is}}) = \sum_a n_a \log(\pi_a), \tag{1}$$

where n_a is the number of times the amino acid a occurred in the alignment. This characterization as ML estimation is useful in suggesting that observed frequencies can be expected to perform well by comparison with full ML estimation when evolutionary distances are large. However, even in the case that evolutionary distances are small, observed frequencies are statistically consistent estimators of stationary frequencies. Indeed, because of site independence, even for a single taxon, the frequency of a given amino acid has the properties of a binomial proportion, and thus, as a consequence of the Law of Large Numbers, converges upon the true frequency with increasing numbers of sites. That likelihood methods can be expected to work even when models are misspecified is a phenomenon that has been exploited in a variety of data settings involving complex dependencies including space-time and longitudinal data modelling (Varin et al. 2011). Log likelihoods like (1) are referred to as composite log likelihoods (Lindsay 1988; Varin et al. 2011). Key properties are that they sum over independent units (sites in the present case) and that while the probability of observing x_i is misspecified (that probability is not $\prod_s \pi_{x_{is}}$ in the present case), the marginal probability of x_{is} ($\pi_{x_{is}}$ in the present case) is correctly modelled (which in the present case amounts to the assumption that frequencies are stationary throughout the tree).

In the case of a frequency mixture the composite likelihood that corresponds to a tree with infinite edge lengths is

$$\sum_i \log(\sum_c w_c \prod_s \pi^{(c)}_{x_{is}}) \tag{2}$$

which we refer to as the multinomial mixture log likelihood. Because the class c for a site is unobserved, the simple reduction in (1) is not applicable and there is no simple formula allowing explicit calculation of the maximizer, $\pi^{(c)}_a$. The EM algorithm of Dempster, Laird and Rubin (1977), however, provides a simple and intuitively appealing scheme to obtain updated frequencies and weights, $\pi^{(cu)}$ and w^u_c, from old ones $\pi^{(c)}$ and w_c. Let n_{ia} denote the number of occurrences of amino acid a at site i. Let $p(c|x_i) \propto w_c \prod_s \pi^{(c)}_{x_{is}}$ denote the conditional probability of class c for a site, under the multinomial mixture model, given the data x at the site; the constant of proportionality is determined by the constraint that $\sum_c p(c|x_i) = 1$. Then updates are obtained through

$$\pi^{(cu)}_a \propto \sum_i p(c|x_i) n_{ia}, \quad w^{(u)}_c \propto \sum_i p(c|x_i) \tag{3}$$

where constants of proportionality are determined from the constraints that $\sum_c w^{(u)}_c = 1$ and $\sum_a \pi^{(cu)}_a = 1$; derivation is given in Supplementary Material. Updating continues until the difference between old and updated parameters becomes small. The updating scheme (3) establishes a relationship between the observed frequencies and those inferred from the mixture model. Had the sites corresponding to class c been known, the observed frequencies would be over those sites. Since they are unknown, estimation takes a weighted average of the frequencies, weighting the contribution from site i more heavily if it was likely from class c.

There are several biases that can be expected to occur with approximate methods like the multinomial mixture approach. We discuss these below and present adjustments.

Penalized Likelihood

Because alignments tend to have few taxa, it is to be expected that some frequencies will be underestimated. With less than 20 taxa, for instance, because there are 20 amino acids, at least one amino acid will not be present at a site. Zero frequencies are likely artefactual in this case. In addition, very small frequencies can cause numerical difficulties in likelihood calculations. An adjustment for small numbers of taxa is to use a penalized log likelihood that adds a penalty term $\eta \sum_c \sum_a \log[\pi_a^{(c)}]$ to the multinomial mixture log likelihood. Here $\eta > 0$ is a tunable parameter that controls the amount of penalization; we investigate a few choices through simulation. Because $\log(\pi_a^{(c)})$ becomes large in magnitude but negative for $\pi_a^{(c)}$ small, adding a penalty term to the multinomial log likelihood prevents frequencies from getting too small. Adding penalization leads to a relatively simple adjustment to the updating scheme (3). Updates of $w_c^{(u)}$ are the same as before but the appropriate update for $\pi_a^{(cu)}$ is shown in Supplementary Material to be $\pi_a^{(cu)} \propto \sum_i p(c|x_i)n_{ia} + \eta$. The approach is consequently comparable to the pseudo-count adjustment discussed, for instance, in Chapter 1 of Durbin et al. (1998). When site classes are known, the pseudo count approach estimates $\pi_a^{(c)}$ as $[\sum_{i \in c} n_{ia} + \eta]/[mN_c + 20\eta]$, where the sum is over the N_c sites that correspond to class

c. Penalized likelihood estimation gives rise to an estimate $[\sum_i p(c|x_i)n_{ia}+\eta]/[m\sum_i p(c|x_i)+20\eta]$, where uncertainty about class membership is adjusted for by weighting sites according to how likely they were to correspond to class c.

Rate Variation Adjustments

Another source of difficulty for multinomial ML is dealing with rate variation across sites. Even with a substantial number of taxa, if a site evolves at a relatively low rate, it is frequently the case that only one or a few amino acids will be present. This is not because the actual frequencies are highly skewed but simply because there is not enough evolutionary distance to observe other amino acids. Since low rate sites are not uncommon, they can give rise to estimated frequency components, $\pi^{(c)}$, that assign mass to one or a few amino acids. Our adjustment is to restrict estimation to sites having rates exceeding the qth percentiles of rates; the choice of q will be investigated. To ensure a range of rates we use the DGPE rate estimates described in (Susko et al. 2003). In brief, DGPE estimates are obtained by fitting a discrete approximation to the gamma rates-across-sites mixture model but with a larger number of components (101 by default). By contrast with the usual approach where rates, $r_k(\alpha)$, depend on the α shape parameter and probabilities, $p_k = 1/K$ are fixed, DGPE fits with rates, r_k, that are fixed and with probabilities, $p_k(\alpha)$ that depend on α. This avoids repeated application of the pruning algorithm and gives an estimate of α. Rates at sites are then estimated on a fixed tree as the conditional means for those sites: $\sum_k r_k p(k|x_i)$ is the rate for site i where $p(k|x)$ is the conditional probability of rate class k as a site, given the data x at that site. The approach gives a larger range of rates at sites than would be estimated under a usual discrete mixture with a few rate classes.

Likelihood Weights for Taxa

The final source of bias that we consider is due to phylogenetic relatedness. Because closely related taxa frequently share the same amino acid at a site, observed frequencies can be dominated by an amino acid shared by a set of closely related taxa. This gives rise to frequency vectors, $\pi^{(c)}$, where the largest $\pi_a^{(c)}$ is overestimated and makes it difficult to estimate frequency vectors that are closer to homogeneous ($\pi_a^{(c)} = 1/20$). Our adjustment for this is a form of likelihood weighting.

To motivate the approach we begin by considering estimation of frequencies at a single site. The weighted composite likelihood is

$$\sum_s v_s \log[\pi_{x_s}] \tag{4}$$

The case $v_s = 1$ corresponds to the usual composite log likelihood that ignores phylogenetic relatedness. We restrict the likelihood weights so that $v_s \geq 0$ and $\sum v_s = m$, the number of taxa; a condition that holds for the usual composite log likelihood. The weights give some flexibility so that closely related taxa can be downweighted and taxa that are more distant from the majority can be upweighted. We implicitly do this by choosing the weights to minimize the variance of the resulting $\hat{\pi}_a$.

The maximizer of (4) is $\hat{\pi}_a = v^T\delta^{(a)}/m$, where $\delta^{(a)}$ has sth component $\delta_s^{(a)} = 1$ if $x_s = a$ and 0 otherwise; see Supplementary Material. It follows that $\text{Var}(\hat{\pi}_a) = v^T\Sigma^{(a)}v/m^2$ where $\Sigma^{(a)}$ is the covariance matrix of $\delta^{(a)}$. That covariance matrix can be calculated (see Supplementary Material) as

$$[\Sigma^{(a)}]_{sj} = \begin{cases} \pi_a(1-\pi_a) & \text{if } s = j \\ p_{aa;sj} - \pi_a^2 & \text{otherwise} \end{cases} \tag{5}$$

Here $p_{aa;sj}$ is the probability of taxa s and j having amino acid a at the site. The average variance of the π_a estimates is then $v^T \Sigma v / m^2$ where $\Sigma = \sum_a \Sigma^{(a)}/20$. Thus to minimize the average variance one needs to minimize $v^T \Sigma v$, subject to the constraint that $\sum_s v_s = m$ and $v_s \geq 0$. An explicit expression for the optimal weights is unavailable but the minimization is a quadratic programming problem and has a global minimizer that can be determined numerically given a Σ. We utilized the R package `quadprog` of Turlack and Weingessel (2013) which implements the methods of Goldfarb and Idnani (1983). Since Σ is unknown in practice it requires estimation. The simple approximation that we use approximates π_a by the observed frequency of amino acid a over all sites and taxa, and $p_{aa;sj}$ by the proportion of sites where both s and j had amino acid a. Alternatively, one might obtain an estimate from the pairwise substitution matrix for s and j where evolutionary distance between the pair is either calculated on a tree or using the pairwise data alone.

To extend the approach above to mixtures of frequencies, we take the composite likelihood contribution at a site for a given class to be weighted: $\prod_s \pi_{x_{is}}^{v_s}$. The resulting weighted composite or multinomial mixture likelihood is then

$$\sum_i \log(\sum_c \prod_s [\pi_{x_{is}}^{(c)}]^{v_s}).$$

This continues to give rise to a simple updating scheme similar to (3) but where n_{ia} is replaced by $\sum_s v_s \delta_{is}$. Allowing for the possibility of penalization, the updates are

$$\pi_a^{cu} \propto \sum_i p(c|x_i) \sum_s v_s \delta_{is} + \eta \quad w_c^{(u)} \propto \sum_i p(c|x_i).$$

In principle the matrix Σ used to obtain likelihood weights should be approximated separately for each class but preliminary experiments suggested the optimal weights were not very sensitive to the stationary frequencies. In all cases we used the simple approximation that approximates π_a by the observed frequency of amino acid a over all sites and taxa, and $p_{aa;sj}$ by the proportion of sites where both s and j had amino acid a.

Tree-Based EM-Updating

The EM-updating scheme given by (3) can be expected to more generally give good estimates of the frequency vectors whenever $p(c|x)$ provides a good approximation to the true $p(c|x)$, the posterior probability calculated using the true generating parameters for the model. Calculating $p(c|x)$ using the multinomial is fast and using weighted composite likelihoods adjust for phylogenetic relatedness to some degree but it is possible that a $p(c|x)$ calculated using a tree will improve upon initial multinomial estimates of the frequency vectors.

EM-updating is generally expected to give good estimates when $p(c|x)$ gives a good approximation to the true $p(c|x)$. To see this, suppose the true $p(c|x)$ is used in (3). Since $n_{ia} = \sum_s I\{x_{is} = a\}$, using the convention that uppercase letters are random, the expected value of an update is

$$\begin{aligned} E[n^{-1}\sum_i p(c|X_i)N_{ia}] &= E[p(c|X_i)\sum_s I\{X_{is} = a\}] \\ &= \sum_s \sum_{x_i|x_{is}=a} \frac{p(x_i|c)w_c}{p(x_i)} p(x_i) \\ &= w_c \sum_s \sum_{x_i|x_{is}=a} p(x_i|c) = w_c \sum_s P(X_{is} = a|c) = w_c m \pi_a^{(c)} \end{aligned}$$

where m is the number of taxa and the sums only consider x_i satisfying that $x_{is} = a$. The final expression is the same as the true frequency $\pi_a^{(c)}$ up to a constant of proportionality that vanishes upon rescaling.

We consider two approaches to obtaining a tree for tree-based EM-updating. One is to calculate a distance matrix for the LG+F model and obtain a neighbour joining (NJ) tree using the neighbour joining method of Saitou and Nei (1987). The second is to calculate a star tree with edge-lengths estimated from the distance matrix through least-squares estimation. The second approach has the advantage that subsequent EM-updates are faster due to there being a single internal node in the tree.

Cross-validation to Estimate the Number of Classes

In many of the simulations we treat the number of mixture classes as fixed and known. In practice, however, they need to be estimated from the data. This cannot be done through multinomial mixture ML estimation since increasing the number of components will always increase the log likelihood; the models are nested. The approach we take is to use cross-validation. The k-fold procedure can be described as follows.

1. Randomly partition the alignment into k separate alignments, $A_1, \ldots, A_k$ of roughly the same size and having no overlap; since n/k might not be an integer the alignment sizes might vary slightly.

2. For each number of classes $C = 2, \ldots, C_{max}$,

    (a) For each predictive alignment A_p, $p = 1, \ldots, k$

        i. Concatenate $A_1, \ldots, A_{p-1}, A_{p+1}, \ldots, A_k$ to create a new alignment $A^{(e)}$. Use this alignment to estimate the frequencies.

        ii. Obtain $l_{C,p} = \sum_{i \in A_p} \log[p_C(x_i)]$, the cross-validated log likelihood, where $p_C(x_i)$ denotes the probability of x under the model but with parameters estimated from $A^{(e)}$.

(b) The cross-validated log likelihood over all folds is calculated as $l_C = \sum_p l_{C,p}$.

Using cross-validated log likelihoods as criteria measures avoids the difficulties associated with models having differing numbers of parameters. Because the data that the log likelihood is being calculated for is completely separate from the data that was used to get estimates, there is little reason to be concerned about the differing numbers of parameters for differing class sizes. There is a caveat to this in that frequency vectors having a number of small entries are likely to be present in both estimated and predictive data sets, and might be predicted via additional classes having small weight. In all applications we considered 10-fold cross-validation with a maximum of 30 frequency classes.

There are two ways of estimating a class from the procedure. The traditional approach is to choose the number of classes, C giving the largest cross-validated l_C (Stone 1977; Smyth 2000). Another, more conservative approach, is to choose C as the first class that gives a larger cross-validated log likelihood than $C + 1$. The conservative approach is motivated by the caveat discussed above whereby excess classes may be estimated as a result of observed frequency vectors having a number of small entries. Similar approaches have been used with other criteria measures in clustering (Gori et al. 2016). Additional motivation for the conservative approach is given in Supplementary Material.

The natural approach to cross-validation is to calculate predicted log likelihoods using the multinomial mixture model. We also consider cross-validated log likelihoods calculated using a NJ tree and using a star tree both constructed as for tree-based EM-updating. Calculation of cross-validated log likelihoods is then more expensive but because parameters are not being estimated under trees, it remains feasible.

Maximum Likelihood Estimation of a Mixture of Frequencies Model on a Fixed Tree

Full ML estimation of frequencies and weights is computationally demanding and dependent on starting frequencies. However, to evaluate how the methods described here compared with ML estimation, we implemented ML estimation using a fixed tree and edge-lengths. The EM algorithm of Dempster, Laird and Rubin (1977) was used. Similarly to multinomial mixture estimation, at each iteration, weight updates are obtained through $w_c^{(u)} \propto \sum_i p(c|x_i)$ but with $p(c|x_i)$ calculated as the posterior probability using the fixed tree and current $\pi^{(c)}$. The updates of the class frequencies at each iteration require numerical optimization. Following the EM scheme, the frequencies for the cth class are obtained by maximizing the contribution to expected complete log likelihood from class c,

$$\sum_i p(c|x_i) \log p(x_i; \zeta, \pi^c).$$

Here ζ includes the tree, edge-lengths and α parameter and are fixed in updating. In simulations and for empirical data we used a NJ tree and DGPE to obtain an α estimate. There is no closed-form expression of the maximizer and so the L-BFGS-B routines of Byrd et al. (1995) and Morales and Nocedal (2011) were required.

Simulation Setting

To evaluate the performance of the multinomial mixture likelihood approach with the adjustments above we consider simulation from a true 21-class mixture model. The classes include the C20 frequency vectors from Le et al. (2008) plus one additional class having the stationary frequencies of the LG model of Le and Gascuel (2008). The frequencies for the

21 classes are given in Figure 1. For a given data set, we generated 1000 sites from each of the 21 classes, using the LG exchangeability matrix and a 4-component discrete gamma rates-across-sites process. The result is a concatenated alignment with 21,000 sites. Data were generated for 74 taxa using the tree given in Figure 2 and $\alpha = 0.74$ for the gamma rate distribution. That generating tree and α were estimated from an expanded version of the Brown et al. (2013) data set with a larger number of taxa.

As a simulation in a more complex setting we also consider estimation for a single simulated data set where each site has its own frequency vector. The tree, edge-lengths and α parameter were the same as above. Frequency vectors at sites were obtained from the posterior mean frequencies at the sites in Brown et al. (2013) data set under a fitted C20+LG+F model. A total of 21,000 site-frequencies were selected at random from the source data for simulation with an LG exchangeability matrix.

Hierarchical clustering of observed frequency vectors over sites provides a default method for frequency estimation. Results for simulated data used the R function `hclust` and average distances between clusters to determine clustering. Because large distance matrices are required, memory constraints can make the approach prohibitive with a large number of sites using implementations like `hclust` that require distance matrices as input. A simpler source of starting frequencies for multinomial mixture ML are provided by the C20 frequencies or other empirical choices. Because the C20 frequencies were the generating frequencies for the simulations, hierarchical frequencies were used for starting values to avoid biasing results. To evaluate how well a set of multinomial mixture ML frequencies did at estimating the true underlying frequencies we calculated the percentage error decrease in L1 distance over hierarchical clustering: $100 \times [L1(h) - L1(m)]/L1(h)$, where $L1(h)$ and $L1(m)$ denote the L1 distances for hierarchical clustering and multinomial mixture ML. For a given set of estimated frequencies, $\hat{\pi}^{(c)}$, the L1 distance is calculated as a sum over amino acids and classes,

$$\sum_{c,a} |\pi_a^{(c)} - \hat{\pi}_a^{(c)}|.$$

A complication arises in that class labeling is arbitrary. For an estimated set of frequencies, $\hat{\pi}^{(c)}$, it is possible, for instance, that the estimated class that best fits the class 1 frequencies of C20 is labeled as class 2. To determine well-fitting classes we used the following scheme.

1. Determine the L1 distance $d_{ck} = \sum_a |\pi_a^{(c)} - \hat{\pi}_a^{(k)}|$ for all pairs of classes c (true) and k (estimated). Continue the following two steps until all classes are matched.

   (i) Determine the two class labels c (true) and k (estimated) giving the smallest d_{ck} among all classes that have not been matched; initially this includes all classes.

   (ii) Class k is the matching class for class c. Remove c (true) and k (estimated) from the set of classes that have not been matched and go to (i).

   Classes are then relabeled so that estimated class c has the same label as the class it matches.

It should be noted that the above approach doesn't guarantee that relabeled classes are best in the sense of minimizing the overall L1 distance between the estimated frequencies. Computing the minimizer through exhaustive search is not feasible.

Empirical Data

We consider 4 empirical data sets listed in Table 1. For each data set, there has been some controversy over the correct topology with different topologies being estimated under mixtures than under a conventional models; see Lartillot et al. (2007) and Wang et al. (2017). Conventional site-homogeneous models differ from frequency profile mixtures in placing *Amborella* as a sister to all other angiosperms for the Amborella data set. For the Microsporidia data set, site-homogeneous models place Microsporidia close to archaea. For the Nematode and Platyhelminth data sets, site-homogeneous models estimate a tree with nematodes or platyhelminths branching at the base of Metazoa, grouping with Fungi to the exclusion of arthropods and deuterostomes. Finally, for the Obazoa data, the position of the breviate protists in the eukaryote tree differs depending on whether a site-homogeneous or frequency profile mixture model is used. Competing trees are given in figures S1-S4 in Supplementary Material. The trees estimated under a conventional, non-mixture model and under the C20 model were included in each case. For the Amborella data, we also calculated log likelihoods for trees previously recovered in Bayesian analyses using the CAT model (Wang et al. 2017). For each tree and data set, edge-lengths and the α parameter were re-estimated via ML estimation.

The JTT exchangeability matrix was found to be the best-fitting matrix for the Amborella data and was used in both non-mixture and mixture fitting. For all other data, the LG exchangeability matrix was used. We adopted the common approach of including a +F component. For default methods, this means that the stationary frequencies of amino acids were determined as the observed frequencies over all taxa and sites. For mixture approaches, the observed frequency vector was used as an additional frequency class, as described in Wang et al. (2014).

Results and Discussion

In what follows we start by considering the extent to which the strategies for estimation described give good estimation of true frequency classes in simulation. Strategies considered include restricting attention to high rate sites in frequency estimation, penalized estimation and using likelihood weights. Starting with unadjusted multinomial ML estimation and considering a sequence of adjustments in turn, results successively lead to set of rough recommendations for frequency class estimation that are utilized in subsequent subsections. For instance, the recommendation to use high rates is used in investigating penalized estimation. Simulation results conclude by considering the effectiveness of cross-validation in estimating the number of classes. Results for empirical data are then considered showing large likelihood increases over default mixture models and good tree estimation.

Restricting Attention to High Rate Sites Improves Estimation

Figure 3 plots the percentage error decrease of multinomial mixture ML (with no penalization or weighting) over hierarchical clustering as a function of q, where only those sites having a rate at or above the qth quantile of rates were used for estimation. It is clearly important to exclude low-rate sites in estimation. Performance is comparable to or even worse than hierarchical clustering if no exclusions are made whereas estimation error decreased by more than 50% over each of the 10 data sets when a quantile threshold of $q = 0.75$ was used; $q = 0.75$ is used in all following analyses.

Inclusion of low rate sites is not problematic when considering a single frequency profile for all of the data. If A has frequency 0.07, for instance, then among sites with a single amino acid, approximately 7% will be A. The difficulty with mixtures is that, since it is unknown which class corresponds to which site, when a large set of sites have a single predominant amino acid (eg. 7% have A) there is a tendency to erroneously group those sites into a single class rather than attributing them to low rates.

The results reported throughout make comparisons with hierarchical clustering. Other clustering approaches are available. As an alternative we compared hierarchical clustering with the popular `kmeans` clustering algorithm (Hartigan and Wong 1979). We considered two starting strategies: (i) choosing the frequency classes over 100 random starting points that give the largest between-to-total sum of squares over 100 random starting points and (ii) using the frequency classes coming from hierarchical clustering as starting points. For both starting strategies, hierarchical clustering was found to perform better. The average percent decrease in error (standard deviation) of hierarchical clustering over `kmeans` was 4.4 (2.1) for starting strategy (i) and 6.1 (2.5) for (ii).

Penalized Estimation Has a Small But Important Effect

Penalized estimation had a relatively small effect on estimation results. Table 2 summarizes results for penalized multinomial ML estimation using high rate sites, no taxa weighting and penalty parameters $\eta = 2$, 5 and 10. The reduction of error over hierarchical clustering was comparable to no penalization, being within 0.5% of the reduction of error without penalization (53.8%); the standard deviations of the reduction were approximately 2%. Clear linear relationships existed between estimated frequencies with penalty to estimated frequencies without penalty ($\eta = 0$). Over all data sets and classes the average R^2 (over data sets) was at least 0.96 over choices of penalty parameter η =2, 5 and 10.

The expectation with penalization is that, for any given class, small frequencies under $\eta = 0$ will be estimated as larger under $\eta > 0$ and large frequencies will consequently decrease. This is supported by fitted regression relationships ($\eta > 0$ frequencies regressed on $\eta = 0$ frequencies) over data sets. The average intercepts suggest an estimated zero frequency with no penalization would be increased to roughly 0.001-0.002 with penalization but the regression slopes, being less than 1, suggest large estimated frequencies with penalization will be reduced.

While the effect of penalization in reducing error is small, it remains valuable as a means of avoiding zero frequencies. With no penalization, over all data sets, classes and amino acids, there were 7 estimated frequencies that were less than 1.0e-8 with $\eta = 0$ but none with $\eta > 0$; one estimated frequency was less than 1.0e-4 with $\eta = 2$. Since it reduces the chance of zero estimates while maintaining comparable performance, we used $\eta = 5$ as a penalty in all following analyses.

Likelihood Weights for Taxa Improves Estimation

Using likelihood weights for taxa gives a substantial improvement over approaches that do not use likelihood weighting. Table 3 shows the reduction in error over hierarchical clustering. Although estimation of frequencies is restricted to high rate sites, it turned out to be valuable to use all of the sites in estimating likelihood weights. Weights using only high rate sites tended to be more homogeneous than those using all sites. Figure 2 gives the average estimated likelihood weights for each of the taxa. The maximum standard deviation in these weights over data sets was 0.33. The weighting is to some degree intuitive. Taxa that are distantly related to most other taxa, and consequently provide less dependent information about frequencies, are upweighted and there is a sampling of relatively large weights throughout the tree.

Figure 1 gives the true frequencies for the frequency classes. Some classes are distinct enough from each other that one can expect that they will be well estimated, but the similarity of many of the frequency classes, suggests difficulties in separating contributions from similar classes. As a measure of how well individual classes were estimated, for each true class and restricting attention to high rate sites for that class, we calculated the average posterior probability for each of the estimated classes; we restricted attention to high rate sites, since it is much more difficult to assign posterior probability to low-rate sites. The results are in Table 4. If an estimated class is highly linked with a particular true class, its average posterior probability will be large only for sites from that true class. We see that this is the case for true classes like Class 5, which exhibits a very unique frequency pattern, but not for Class 21 which has more homogeneous frequencies that are similar to those of a number of other classes.

While large improvements over hierarchical clustering have been found, improvements can be expected to be less when fewer sites are considered. Table S1 in Supplementary Material considers frequency estimation using the same approach as in Table 4 and the same simulation setting but with only 500 sites and with different numbers of classes; separate simulations were conducted using the first $C = 5$, 10, 20 and the full 21 classes in Figure 1. Average percent decreases were smaller than the 71.2% reported in Table 4. With only 500 sites, decreases were in the range 8.9%-18.1% with larger standard deviations over the 10 simulated data sets.

Tree-Based EM-Updating Improves Estimation with a Small Number of Iterations

Table 5 gives the results of tree-based EM-updating on a full phylogenetic tree applied with starting frequencies coming from the best performing method in Table 3; the average error, $|\hat{\pi}_a^{(c)} - \pi_a^{(c)}|$, in estimation is reported in Supplementary Material Table S1. For all approaches, performance improves in initial iterations but then remains relatively stable or decreases. Similarly, log-likelihoods after updating, increased most substantially from 1 to 10 iterations and then became more stable (Supplementary Material Figure S6). Since error decreases were best with 5 or 10 iterations, one possible stopping strategy for updating is to stop when log likelihoods on the tree show relatively small increases. In practice the true tree is unknown, but performance with it provides an upper bound on what may be achievable with good estimates of the tree. Using an estimated NJ tree gives comparable performance.

Part of the reason for the lack of improvement in performance as the number of iterations increases, which occurred across methods, likely has to do with numerical instabilities due to some frequencies being estimated as close to 0. Due to the penalized likelihood estimation used to obtain the starting frequencies, for each data set, the minimum starting frequency over all classes and amino acids was at least 2×10^{-4}. This minimum decreased over all data sets as the number of iterations increased. Using the true tree, after 50 iterations the minimum frequency ranged between 5×10^{-5} and 5×10^{-11}. Since the true tree adjusts for phylogenetic relatedness, the small frequencies are primarily a consequence of using a relatively small number of taxa. In the case that the number of taxa are small, if an amino acid, a, has low frequency for a particular true frequency class, c', then a might not be observed ($n_{ia} = 0$) at sites i corresponding to c'. If in addition, the frequencies for this true class differ substantially from those of other classes, then it will be easy to distinguish site patterns that correspond to c' from those of other classes. Consequently, $p(c^*|x_i)$ will be large for the estimated class, c^*, that best fits c' if, and only if, x_i corresponds to true class c'. Since n_{ia} is small or zero at such sites, the weighted average, $\sum_i p(c^*|x_i)n_{ia}$, used to update the frequency of a in (3), will be small.

A variation on the explanation above is also important in understanding why updating mixture weights lead to performance decreases with increasing numbers of iterations. The added flexibility of updating weights often leads to mixture weights that are small with corresponding frequency classes that are dominated by a few frequencies. With 50 iterations, some data sets had classes that, up to machine precision, had frequencies for some amino acids equal to 0.

A difficulty with tree-based EM-updating is that it comes with a substantial computational cost due to the need to repeatedly calculate likelihoods on trees. Using a star tree reduces this computational cost substantially. No attempt was made to optimize the software used and results will vary depending on hardware but, using one particular data set for illustration, the elapsed (wall clock) time required for 50 posterior updates was approximately 5.5 hours using an NJ tree and 10 minutes using a star tree. There was however, a small performance decrease due to using the star tree.

Cross-validation Requires Log Likelihoods Calculated on a Tree

Figure 4 gives plots of the cross-validated log likelihoods for estimation of the number of classes in the mixture; the number of classes in the simulating model is $C = 20$. Regardless of how the log likelihood is calculated, the initial rate of increase, as a function of the number of classes, is large. There is clear evidence that smaller numbers of classes are not sufficient. Using the multinomial log likelihood for fitting does not work well at estimating a sufficient number of classes. The cross-validated log likelihood increases steadily. Some of this may be due to the presence of low rate sites in the test samples. Adding classes allows for frequency vectors that are large for a few amino acids and that fit low rate sites well. Adjusting for rate variation is thus important which is why using the star tree or NJ tree to estimate the number of classes gives cross-validated likelihood curves that become relatively flat with larger numbers of classes. The star tree, however, still gives log likelihoods that increase too quickly. As Table 6 indicates, for 3 of the simulated data sets, the cross-validated log likelihood was maximized using 29 classes, but for the other 7 data sets, it continued to increase over all numbers of classes. Using the NJ tree gave much better performance but tended to over-estimate the number of classes.

An alternative approach to estimation of the number of classes via cross-validated log likelihoods is to choose the first class, C, which has a larger cross-validated log likelihood than $C+1$. Using this approach tended to gives smaller numbers of estimated classes regardless of the way in which cross-validated log likelihoods were calculated. The star tree still tended to over-estimate whereas the NJ tree tended to under-estimate the number of classes. Because many of the frequency classes were similar to each other (Figure 1) under-estimation of the number of classes might not cause difficulties for downstream analyses.

That estimation under a tree is needed for cross-validation is further illustrated in Figure S7 of Supplementary Material where simulation is from a single frequency vector ($C = 1$). Cross-validated estimation with a star or NJ tree give the correct number of components whereas the cross-validated multinomial log likelihood is increasing as a function of C.

Maximum Likelihood Estimation on a Tree is Expected to Improve Performance

with Good Tree Estimates

Table 7 gives the results of full ML estimation on a fixed tree and with fixed edge-lengths, applied with starting frequencies coming from the best performing method in Table 3. Using the NJ tree gives performance improvements over the starting frequencies but the improvements are comparable to those using posterior weighting (Table 5). No attempt has been made to optimize software but, using one particular data set for illustration, the elapsed (wall clock) time required for 50 posterior updates was approximately 5.5 hours whereas more than 8 days were required for ML estimation.

Maximum likelihood estimation using the true tree is not possible in practice but performance with it provides an upper bound on what is achievable with good estimates of the tree. Comparing results with those of tree-based EM-updating on the true tree (Table 5), a substantial improvement is obtained and the approach gave the best percentage error decrease of any method considered.

Large Variability of Frequency Vectors Makes Estimation Harder

The final simulation result we briefly consider is for the setting were each site has a completely different frequency vector. To ease computation multinomial ML estimation was used with no EM-updating; high rate sites, likelihood weights and penalization continued to be used and a +F component was included. Results were compared with C20+F frequency vectors and from hierarchical clustering.

The log likelihoods were largest for frequencies estimated using multinomial ML (With 20 classes +F $\triangle$LnL=269.6 by comparison with hierarchical clustering and $\triangle$LnL=81.4 by comparison with C20+F) suggesting it gave the best fit. Because each site has its own frequency vector, it is not longer possible to evaluate the abilities of the methods to estimate the class frequency vectors. To compare methods we considered the average difference between frequencies at a site and the posterior mean frequency estimate using the frequencies for 20 class and 60 class models but with different frequency estimates. All choices of frequencies gave similar errors in estimation of approximately 3.0 with comparable standard deviations of approximately 4.0.

Large Likelihood Increases are Obtained with Empirical Data

We obtained frequency vectors using multinomial mixture estimation and tree-based EM-updating for the four empirical data sets listed in Table 1. Due to the good performance found in simulations, for each data set, multinomial mixture estimation was applied to sites with rates larger than the 75th percentile, using estimated likelihood weights and penalized estimation with $\eta = 5$. Tree-based EM-updating was conducted using the estimated NJ tree and a gamma rates-across-sites distribution with 4 rate categories and α estimated using DGPE. In each case models were fit with $10, 20, \ldots, 60$ classes with starting frequencies coming from the C-series frequency classes.

Figure 5 gives the log likelihood increases for fixed trees when frequencies used in likelihood calculation were estimated using multinomial mixture ML estimation and tree-based EM-updating; increases are over the likelihoods for the C-series model with the same number of components. Regardless of the tree or data set considered, enormous gains in likelihood were obtained. The smallest likelihood increase over all data sets, trees and methods is 1649.4. The C-series models are nested within the mixture models, and the mixture model has $380 = 19 \times 20$ additional parameters. Using likelihood theory, if the C20+F+Γ model were correct, the chance of observing a likelihood increase as large as 1649.4 is approximately $P(\chi^2_{19C} > 2 \times 1649.4)$, which up to machine precision is 0, for any $C = 10, 20, \ldots, 60$. In most cases, the log likelihood increase gets larger with larger C, with the Amborella and Platyhelminths data sets providing exceptions when $C = 60$. Using likelihoods as a measure of fit, tree-based EM-updating tended to give substantially larger likelihood increases than using multinomial frequencies; the Obazoa data provided an exception, however.

Figure 6 gives the log likelihood differences between the mixture tree over the default tree when frequency classes are obtained using multinomial ML estimation or EM-updating. The relatively small increases in Figure 6 may be a bit surprising because of the enormous increases in likelihood using the new approaches over C-series models fitted to a fixed tree (Figure 5). In each case considered in Figure 6, a positive log likelihood difference implies that the mixture tree was favoured over the true tree. With the exceptions of some settings for the Platyhelminths and Microsporidia data, the mixture tree is always favoured.

By contrast with simulated data, starting frequencies came from the C-series frequency classes. Since C-series generating frequencies were used in simulation, the intent in simulations was to avoid bias in making comparisons to hierarchical clustering. Because the behaviour of hierarchical clustering is not clear in more complex real data settings, fixed C-series frequency classes may be considered preferable. Figures S7-S8 in Supplementary Material give results for multinomial mixture ML estimation like those of Figures 5-6 but using starting frequencies coming from hierarchical clustering; due to the large memory requirements, the R package `Rclusterpp` (Linderman and Bruggner 2013) was used in place of the default clustering algorithm `hclust`. While different starting points often give different solutions, the general trends and conclusions are the same as for Figures 5-6.

For the Microsporidia, Nematode and Obazoa data the estimated C20+F+Γ tree usually gave the largest likelihood for the mixture models. The support for the correctness of the Obazoa tree is primarily through mixture model-based analyses (Brown et al. 2013) whereas additional support (Keeling and Fast 2002, Brinkmann et al. 2005) has been given to microsporidia+fungi grouping in the C20+F+Γ tree. For the Nematode data, the C20+F+Γ Ecdysozoa tree is also not controversial. In addition to being supported by mixtures, Lartillot et al. (2007) show that the Ecdysozoa tree (nematodes+arthropods) is obtained with conventional models when closer outgroups to bilaterian metazoans are included in the data set; specifically, choanoflagellates and a cnidarian.

The Platyhelminths data consider the same proteins and sites as in the Nematode data. Many of the taxa are the same but the Platyhelminths dataset replaces the data for the 10 nematodes with sequences from 5 platyhelminths. The CAT+GTR tree for the platyhelminths data differs from the LG+F+Γ tree in supporting a Protostomia grouping (platyhelminths + arthropods to the exclusion of deuterostomes) rather than a Coelomata grouping (deuterostomes + arthropods to the exclusion of platyhelminths). That tree was also obtained using the PMSF methodology of Wang et al. (2017). However, the C-series mixtures estimate the incorrect Coelomata tree. While the log likelihood difference in favour of the Coelomata tree gets smaller for C-series mixtures as C increases, the Coelomata topology continues to be estimated. Using multinomial and updated mixture frequencies in place of the C-series frequencies, the correct Protostomia position is favoured with $C \geq 20$ for the multinomial frequencies and $C \geq 30$ with tree-based EM-updating.

For the Amborella dataset, the main difference between trees, whether *Amborella* is at the base of the angiosperms (JTT+C20 tree) or forms a clade with the water lilies, remains contentious (Leebens-Mack et al. 2005; Wickett et al. 2014; Drew et al. 2014; Goremykin et al. 2015; Rokas et al. 2017). Perhaps not surprisingly then, the pattern of likelihood increases differed for the Amborella dataset by comparison with the other data sets. Whereas multinomial and updated mixture frequencies usually showed comparable or larger increases for mixture-derived trees over default estimation, for the Amborella dataset, the log likelihood increases tended to be smaller and decreased with increasing C. Nevertheless, the C20+F+Γ and CAT+GTR trees tend to be favoured. Interestingly, whereas the C-series models consistently favoured the C20+F+Γ tree over the CAT+GTR tree, which differ only in their placement of *Calycanthus*, the multinomial and updated mixture frequencies gave much more comparable likelihoods for the two trees.

Figure 7 gives the results of cross-validation for the multinomial mixture ML frequencies. By contrast with the simulated data, the additional frequency variability for the empirical data suggests a large number of components. The largest cross-validated log likelihood for the range of C considered is always at $C = 60$. For the Platyhelminth data, the NJ-Tree cross-validated log likelihood for $C = 20$ is larger than for $C = 30$ and the star tree cross-validated log likelihood increases slowly going from $C = 20$ to $C = 30$. Similarly for the related Nematode data, the NJ-Tree increase from $C = 20$ to 30 is relatively small, suggesting that for these data sets, $C = 20$ might give a reasonable choice with limited computational resources. The biggest increase in cross-validated log likelihood for all methods is from $C = 10$ to 20. Thus cross-validation always supports least 20 classes; the only choice of C that sometimes did not support the mixture tree with multinomial mixture ML frequencies was $C = 10$ (Figure 6).

Because frequency classes with a few relatively large frequencies will give rise to fewer distinct amino acids at a site, it is possible that they can be explained to some degree by allowing a richer rates-across-sites distribution. We expect, however, that richer rates-across-sites will not completely explain frequency classes with a few relatively large frequencies because sites corresponding to such classes are expected not only to show fewer amino acids but fewer amino acids of particular type; eg. consistently R and K for Class 4 in 1. To test this hypothesis we fit much richer, effectively unconstrained, rate distributions referred to as the discrete estimate (DE) in Susko et al. (2003). The mixtures allow any distribution on a set of 100 rates logarithmically equal-spaced from 0.01 to 10, plus a zero rate. If it is true that richer rates-across-sites distributions do not provide a sufficient explanation, it is expected that log likelihood increases of a frequency mixture model over the single frequency will continue to increase substantially as a function of C when DE is used as the rate distribution. This is indeed the behaviour exhibited in Supplementary Material Figure S10. The rapid increases due to richer mixture of frequency models is the predominant feature of Figure S10, suggesting that frequency variation was the more important model element. However, the increases in log likelihood over the single frequency model were usually smaller using DE than gamma rate variation and the rate of increase was usually slower. Thus the frequency mixtures seem to be explaining some of what may be rate variation and/or richer rate models help to explain frequency variation to some degree. Further evidence that the rate variation is entangled with frequency variation is provided by Supplementary Material Figure S11 which gives the estimated cumulative distribution functions of rates. These usually are closer to continuous (fewer plateaus) when fitted with a single frequency class.

Conclusions

Multinomial mixture ML estimation showed good performance in computational experiments provided that several adjustments were made to the base methodology. It is important to restrict use to sites with relatively high rates. While we found that restricting attention to the top quartile of rates gave optimal performance, optimal rate thresholds may vary depending on the nature of the data. Likelihood weighting helps and using penalized estimation can prevent frequencies from getting too small. Tree-based EM-updating was found to sometimes provide further performance improvements albeit with additional computational cost.

Several additional adjustments to the approaches might be considered. One is to use cross-validation to estimate the penalty parameter η in penalized approaches. This increases computational costs and was not pursued here due to the similarity of frequencies with different choices of η and due to the modest performance gains.

Most of the adjustments to the base methodology were motivated by difficulties estimating vectors of 20 frequencies at a site using the limited information provided by relatively small number of dependent taxa. Performance with very large numbers of well-separated taxa might not require as many adjustments. For the commonly occurring case that there are less than 100 taxa, however, multinomial mixture ML estimation provides a computationally feasible means of estimating mixture profiles that can be expected to give good performance.

Acknowledgments

This research was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada. The authors thank Huaichun Wang for valuable advice and help with the simulation settings and empirical data analyses.

References

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.

Byrd, R.H., Lu, P., Nocedal, J., Zhu, C. 1995. A limited memory algorithm for bound constrained optimization. SIAM J. Sci. Comput. 16, 5, 1190-1208.

Dempster, A.P., Laird, N.M., Rubin, D.B. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. J. Royal Statist. Soc., Series B. 39:1-38.

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.

Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998). Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge University Press.

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

Goldfarb, D., Idnani, A. 1983. A numerically stable dual method for solving strictly convex quadratic programs. Math. Programming 27:133.

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.

Gori, K., Suchan, T., Alvarez, N., Goldman, N., Dessimoz, C. 2016. Clustering genes of common evolutionary histories. Mol. Biol. Evol. 33:1590–1605.

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

Hartigan, J.A. and Wong, M.A. 1979. A K-means clustering algorithm. Appl. Statist. 28:100–108.

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.

Keeling, P.J., Fast N.M. 2002. Microsporidia: Biology and evolution of highly reduced intracellular parasites. Annu. Rev. Microbiol. 56:93116.

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. 29:16951701.

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.

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(Suppl 1):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.

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.

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.

Linderman, M. and Burggner, R. (2013). Rclusterpp: Linkable C++ clustering. R package version 0.2.3.

Lindsay, B.G. 1988. Composite Likelihood Methods. Contemporary Mathematics, 80:221–239.

Morales, J.L. and Nocedal, J. 2011. Remark on "Algorithm 778: L-BFGS-B: Fortran Subroutines for Large-Scaled Bound Constrained Optimization" ACM Trans. Math. Soft. 38, No. 1. Article 7.

Philippe H., Brinkmann H., Lavrov D.V., Littlewood D.T.J., Manuel M., Wrheide 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., Wrheide 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.

Saitou, N., Nei, M. 1987. The neighbor-joining method: A new method for reconstructing evolutionary trees. Mol. Biol. Evol. 4:406–425.

Shen, X., Hittinger, C.T., Rokas, A. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nature Ecol. Evol. 1:0126.

Smyth, P. 2000. Model selection for probabilistic clustering using cross-validated likelihood. Statistics and Computing. 9:63–72.

Stone, M. 1977. An asymptotic equivalence of choice of model by cross-validation and Aikaike's criterion. J. Royal. Statist. Soc. Series B. 39:44–47.

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.

Turlach, B.A. and Weingessel, A. 2013. quadprog: Functions to solve quadratic programming problems. R package version 1.5-5.

Varin, C., Reid, N. Firth, D. 2011. An overview of composite likelihood methods. Statistica Sinica 21:5–42

Wang, H., Minh, B. Susko, E., Roger, A.J. (2017). Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. To appear in Syst. Biol.

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. 31:779792.

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.

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.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39:306–314.

Table 1: Empirical data sets.

Dataset		Proteins	Taxa	Sites	Source
1	Amborella	61	24	15688	Leebens-Mack et al. (2005)
2	Microsporidia	133	40	24291	Brinkmann et al. (2005)
3	Nematode	146	37	35371	Lartillot et al. (2007)
4	Platyhelminths	146	32	35371	Lartillot et al. (2007)
4	Obazoa	159	68	43615	Brown et al. (2013)

Table 2: Summary of results for penalized estimation. The first row gives the average percent reduction of error (standard deviation) over hierarchical clustering for different penalty parameters. The second row gives the average R^2 for the regression of frequencies with penalization on frequencies without penalization ($\eta = 0$); class labels with penalization were chosen to best match frequencies without penalization. The third and fourth rows give the intercept and slope of the regressions.

	Penalty Parameter η			
	0	2	5	10
Percent Decrease	53.8 (1.8)	53.3 (2.4)	53.8 (2.4)	54.0 (2.3)
R^2		0.98 (0.04)	0.98 (0.04)	0.96 (0.04)
Intercept × 100		0.09 (0.13)	0.11 (0.13)	0.20 (0.12)
Coefficient		0.98 (0.03)	0.98 (0.03)	0.96 (0.02)

Table 3: The average percent error decrease of multinomial mixture ML over hierarchical clustering for simulated data when likelihood weights for taxa are used. Weights were approximated from the entire data set as well as for only those sites with rates larger than the 75th percentile of rates.

Method	Error Decrease	SD
Optimal Weights for Entire Data	71.2	1.8
Weights Optimized for High Rate Sites	64.8	3.8
No Weights	53.8	2.4

Table 4: For each true class and restricting attention to high rate sites for that class, the average posterior for the esimated classes. Top-ranked estimated classes are listed with, in parenthesis, the average posterior probability. Results are for one of the simulated data sets and posteriors were calculated using the true tree and edge-lengths.

True Class	Estimated Classes (Posteriors)	True Class	Estimated Classes (Posteriors)
1	6 (0.96)	14	7 (0.94)
2	12 (0.93)	3	15 (0.92)
12	2 (0.87) 8 (0.10)	13	8 (0.33) 15 (0.33) 19 (0.22) 12 (0.08)
4	13 (0.82) 10 (0.16)	5	21 (0.95)
6	18 (0.73) 1 (0.20)	20	4 (0.49) 18 (0.42)
7	16 (0.96)	10	17 (0.97)
16	3 (0.98)	17	5 (0.95)
8	20 (0.79) 14 (0.10) 8 (0.05)	9	8 (0.54) 19 (0.43)
11	10 (0.46) 1 (0.42) 19 (0.07)	15	11 (0.96)
18	19 (0.89) 1 (0.05)	19	9 (0.65) 19 (0.14) 10 (0.11)
21	19 (0.42) 20 (0.31) 8 (0.11) 1 (0.08)		

Table 5: The average percent error decrease (standard deviation) over hierarchical clustering for simulated data when tree-based EM-updating is used with a fixed number of iterations. Tree-based EM-updating uses (Weight Updating): an estimated neighbour-joining tree and updating of mixture weights, (Star Tree): an estimated star tree with likelihood weights but no updating of mixture weights (NJ Tree): an estimated neighbour-joining tree with likelihood weights but no updating of mixture weights, and (True Tree): the true tree, edge-lengths and mixture weights with no likelihood weighting.

Iteration	Weight Update	Star Tree	NJ Tree	True Tree
0	71.2 (1.8)	71.2 (1.8)	71.2 (1.8)	71.2 (1.8)
1	78.7 (1.8)	78.8 (1.8)	78.7 (1.7)	80.9 (1.7)
5	80.2 (3.1)	81.2 (2.9)	83.7 (3.0)	85.0 (2.2)
10	77.6 (4.0)	81.0 (3.3)	84.1 (3.8)	84.2 (2.7)
25	69.4 (4.6)	79.2 (4.4)	83.1 (4.7)	82.8 (3.7)
50	57.9 (6.5)	76.5 (4.6)	81.5 (5.3)	82.5 (4.0)

Table 6: The estimated number of classes using either the estimated NJ tree or star tree to calculate log likelihoods. The estimated class was chosen either to maximize the cross-validated log likelihood (LnL) or as the first class C such that the cross-validated log likelihood for C was large than the cross-validated log likelihood for $C + 1$ ($\triangle$LnL). The number of classes in the simulating model is $C = 20$.

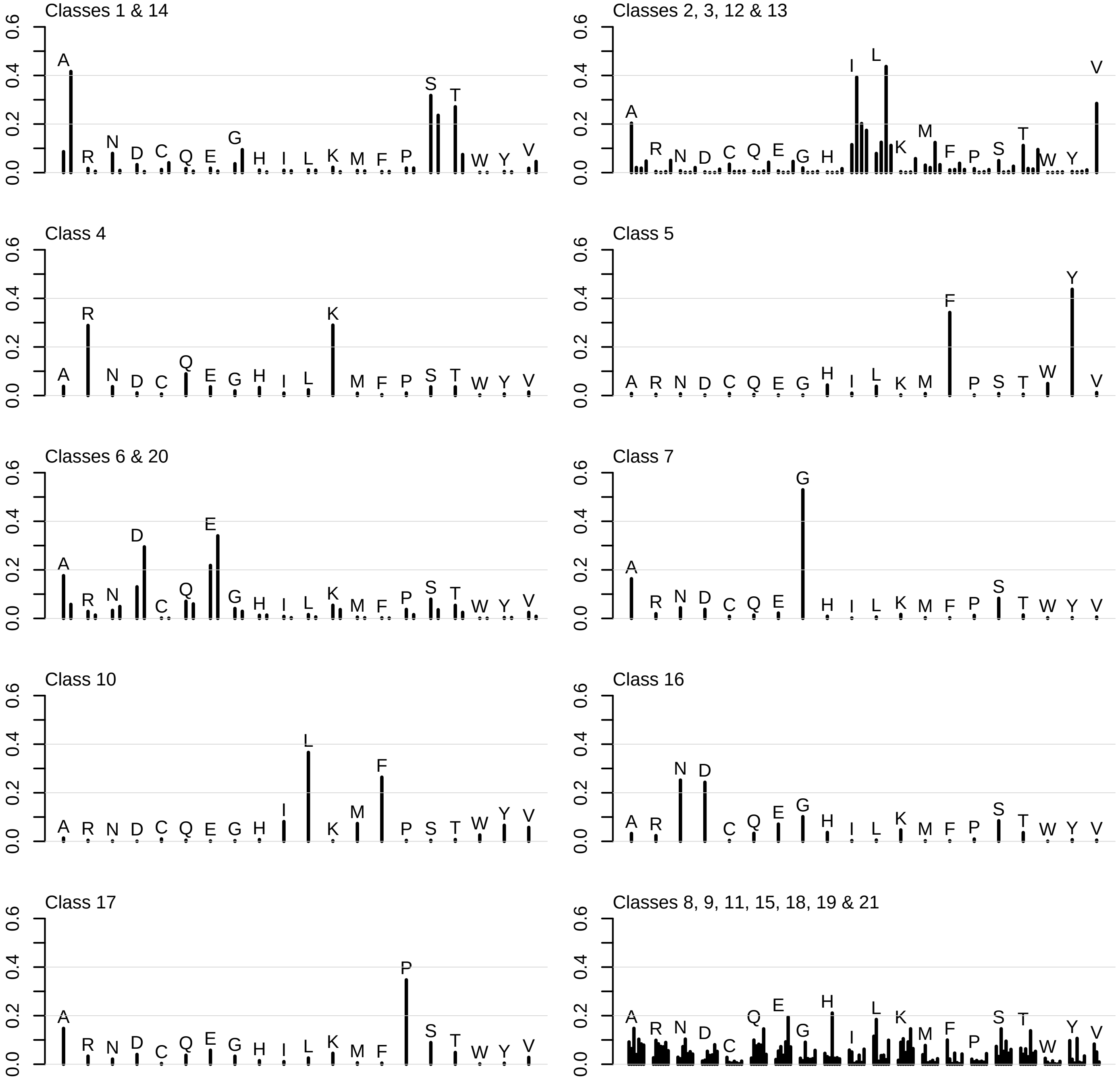
NJ (LnL)						
Number of Classes	20	21	22	24	25	
Number of Datasets	3	1	2	3	1	
NJ ($\triangle$LnL)						
Number of Classes	16	17	19	20		
Number of Datasets	1	3	2	4		
Star (LnL)						
Number of Classes	29	30				
Number of Datasets	3	7				
Star ($\triangle$LnL)						
Number of Classes	20	21	22	24	25	27
Number of Datasets	3	1	3	1	1	1

Table 7: The average percent error decrease (standard deviation) over hierarchical clustering for simulated data after full ML estimation of frequencies and weights after a fixed number of iterations and using a fixed tree. For comparison, some results from Table 5 corresponding to tree-based EM-updating are repeated.

	Updating		ML Estimation	
Iteration	NJ Tree	True Tree	NJ Tree	True Tree
0	71.2 (1.8)	71.2 (1.8)	71.2 (1.8)	71.2 (1.8)
1	78.7 (1.7)	80.9 (1.7)	79.8 (1.6)	82.9 (1.5)
5	83.7 (3.0)	85.0 (2.2)	79.6 (1.5)	88.9 (1.2)
10	84.1 (3.8)	84.2 (2.7)	79.9 (1.3)	91.3 (2.0)
25	83.1 (4.7)	82.8 (3.7)	80.7 (3.2)	92.5 (4.1)
50	81.5 (5.3)	82.5 (4.0)	80.6 (3.3)	93.1 (4.3)

Table 8: The average error in estimation, $|\pi_a^{(c)} - \hat{\pi}_a^{(c)}|$, multiplied by 100 (standard deviation) and average percent error decrease (standard deviation) over hierarchical clustering for simulated data with $n = 500$ sites and $n = 21,000$ sites.

	Sequence Length	
	500	21,000
Average Error	4.16 (0.17)	1.38 (0.08)
Percent Decrease	15.8 (2.3)	71.2 (1.8)


Classes 1 & 14

Classes 2, 3, 12 & 13

Class 4

Class 5

Classes 6 & 20

Class 7

Class 10

Class 16

Class 17

Classes 8, 9, 11, 15, 18, 19 & 21

Figure 1: The frequencies for the 21 classes used to simulate data. Similar frequency classes have been grouped together.

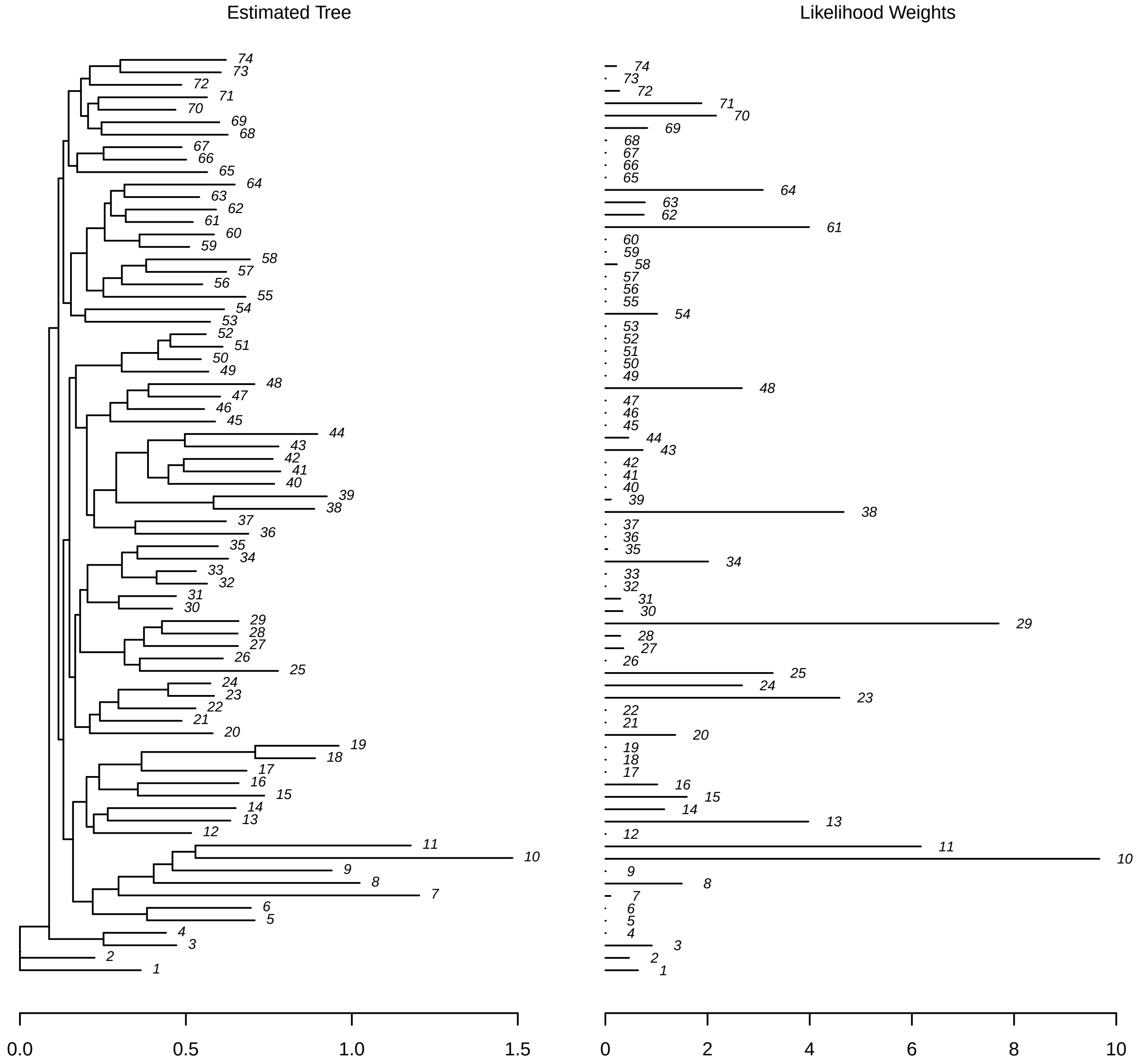


Figure 2: The tree used to simulate the data sets and the average estimated optimal likelihood weights for the tree over data sets.

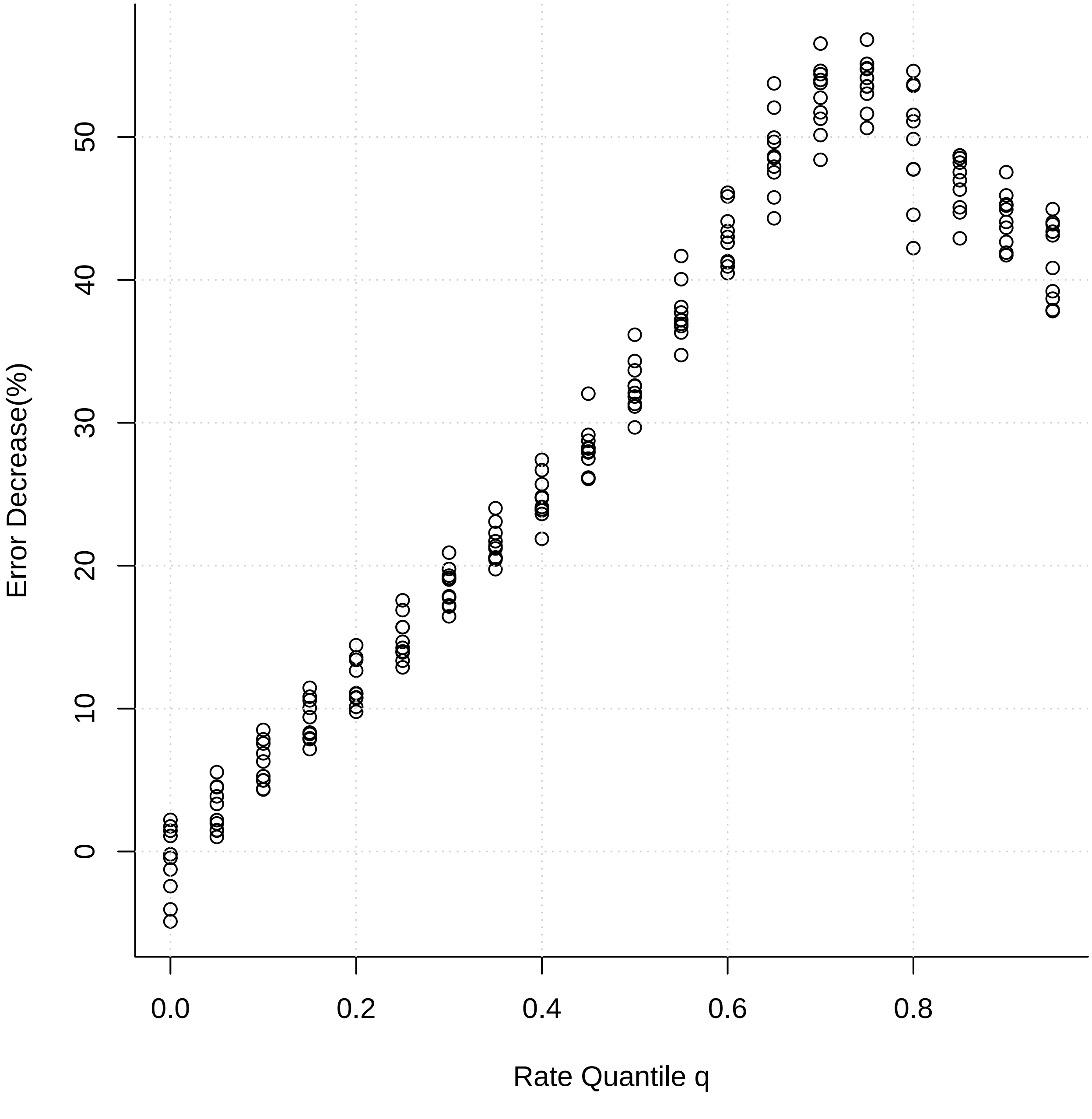


Figure 3: Error decrease of multinomial mixture ML over hierarchical clustering for simulated data when sites with rates at or above the qth quantile of rates are used in estimation. At each values of q, performance was evaluated over the same 10 simulated data sets.

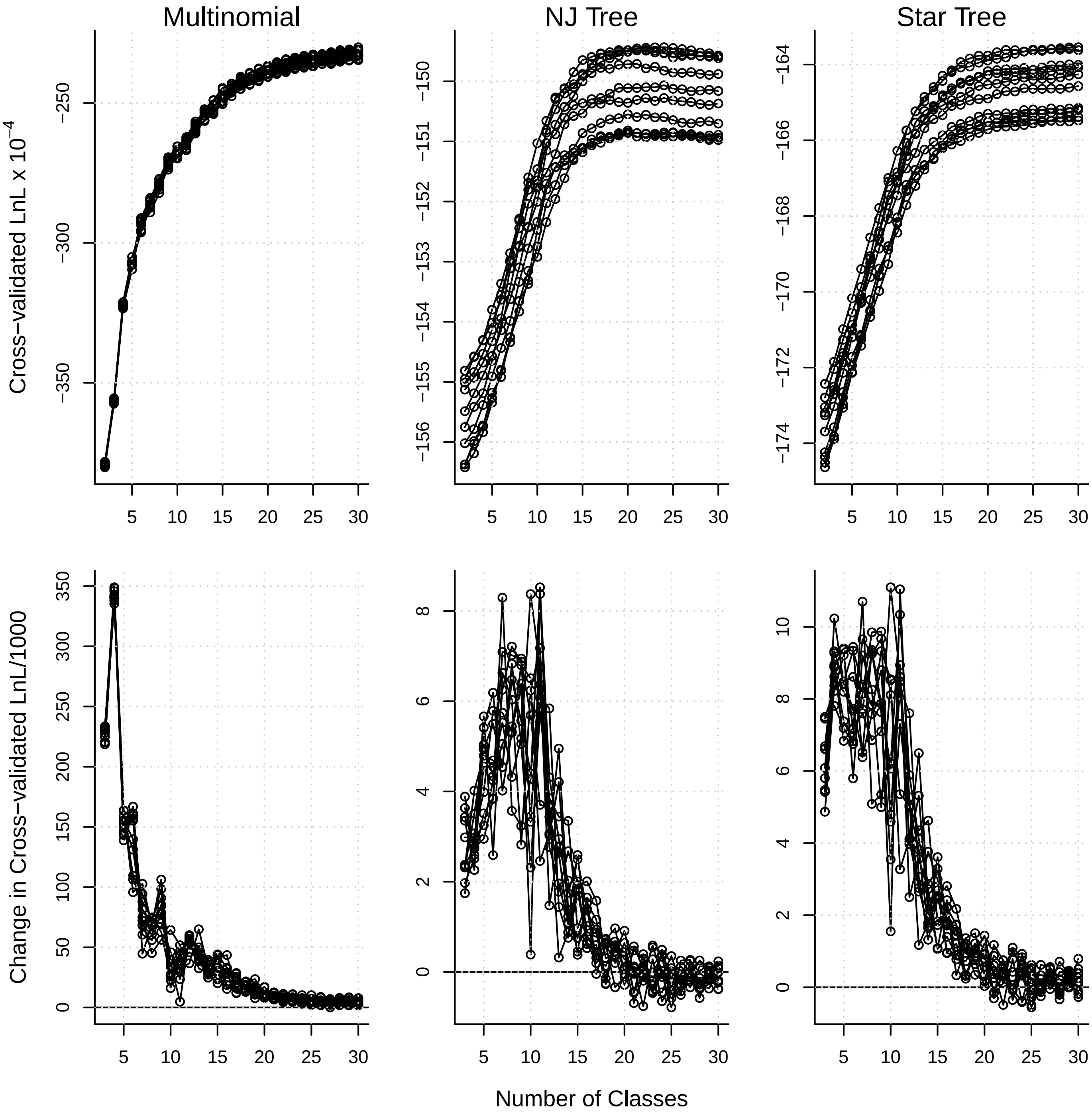


Figure 4: For each of the 10 simulated data sets, the cross-validated log likelihoods and changes in cross-validated log likelihoods (LnL for class C+1 - LnL for class C) as the number of classes, C, increases. Log likelihoods were calculated under the multinomial mixture, using the NJ tree and an estimated star tree. The number of classes in the simulating model is $C = 20$.

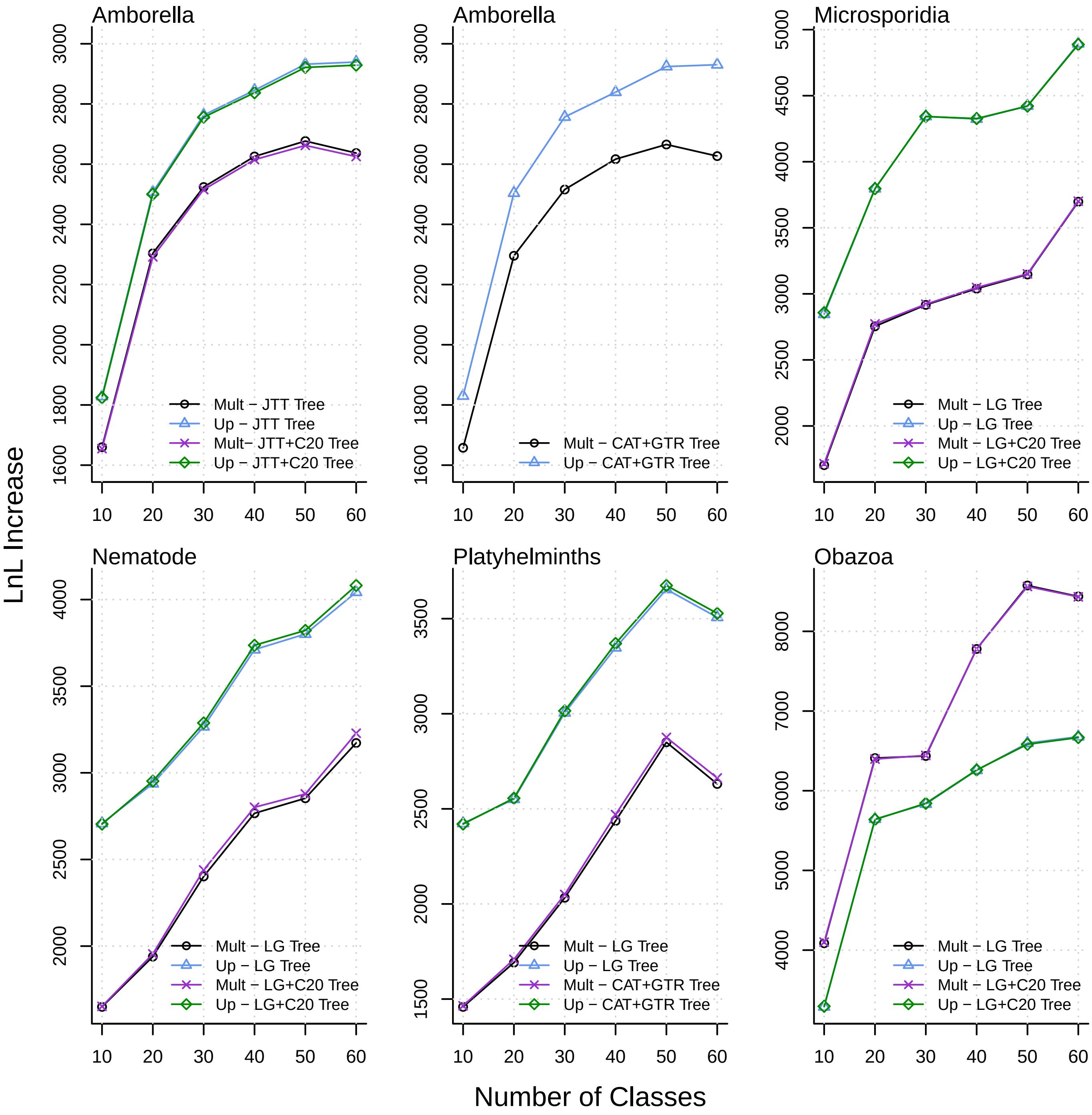


Figure 5: Increases in log likelihoods for fixed trees when frequencies used in likelihood calculation were estimated using multinomial mixture ML (Mult) and tree-based EM-updating (Up). Each increase is the difference in log likelihood over that of the C-series model with the same number of classes. Models used in tree-estimation and likelihood evaluations always included a +F+Γ component.

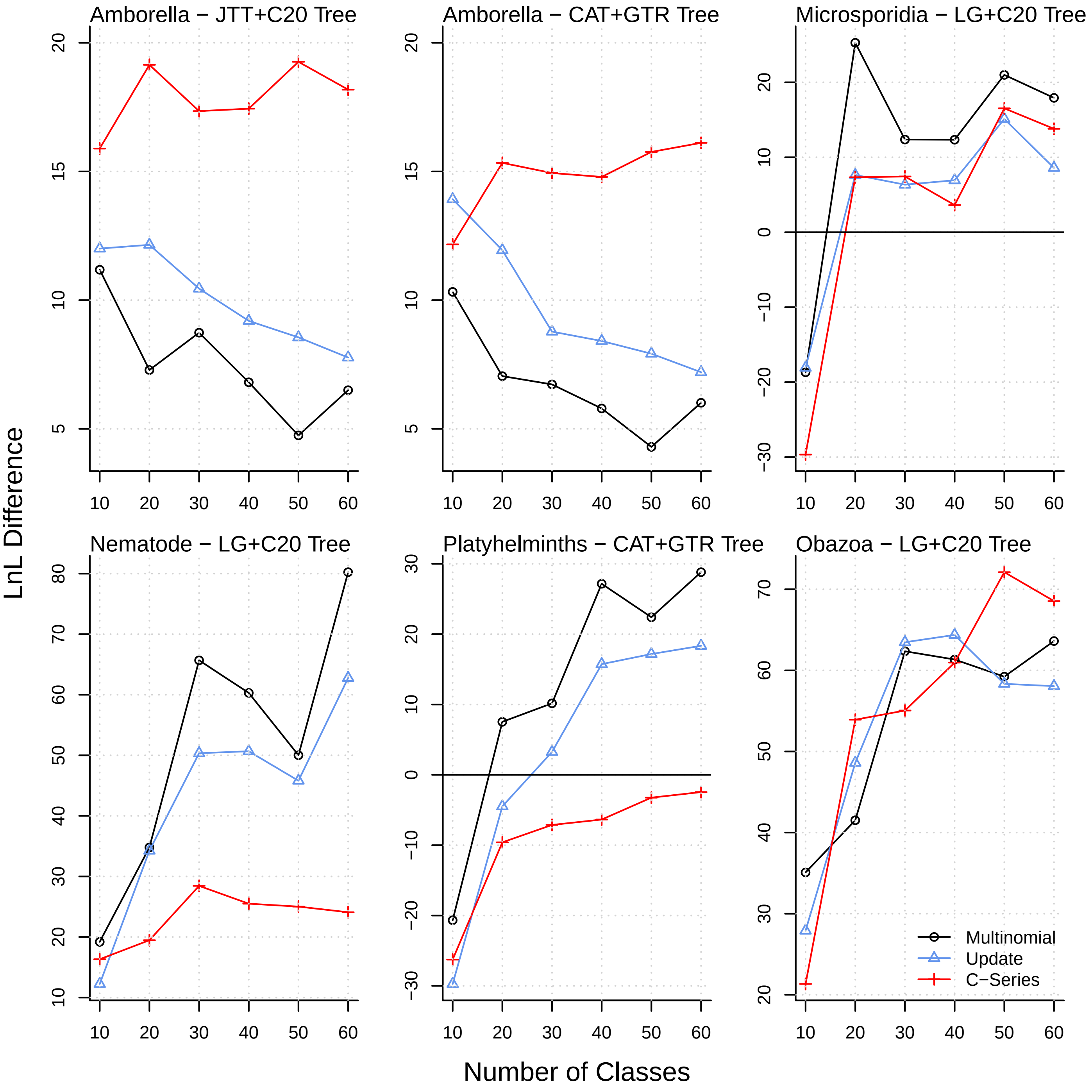


Figure 6: Log likelihood differences for mixture trees over trees estimated using default models that do not allow mixtures of frequencies. Models used in tree estimation always included a +F+Γ component. Values above the indicated $y = 0$ line imply that the mixture tree was favoured by the approach.

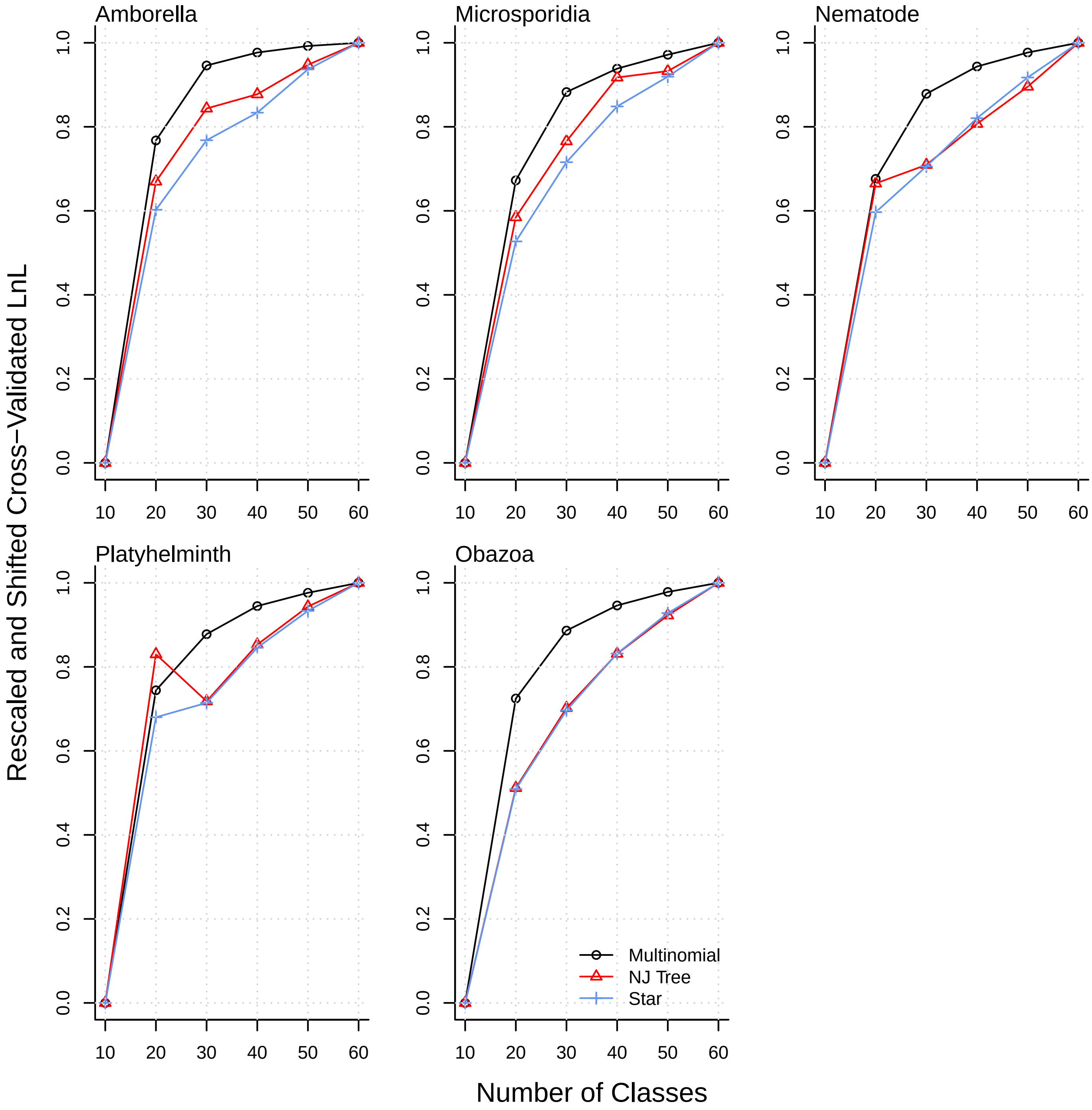


Figure 7: Rescaled and shifted cross-validated log likelihoods for the multinomial mixture ML frequencies. To allow comparisons across different likelihood calculations, cross-validated log likelihoods have been rescaled and centered to have minimum 0 and maximum 1.