Comparison of finite and infinite mixture models for capturing compositional heterogeneity across sites

by

Thomas Bujaki

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of

Master of Science

in

Carleton University

Ottawa, Ontario

Acknowledgements

I would like to thank Dr. Rodrigue and Dr. Avis for being my co-supervisors.

I would like to thank Hao Wang and Omar Kazmi for helping me understand and work with the statistics and computer science aspects of phylogenetics research.

I would like to thank Emma Groulx for her tremendous emotional support over the course of this Master's.

Finally, I would like to thank my family and friends, especially my parents, Merridee and Peter Bujaki for their financial support and wise council throughout my education.

Abstract

Phylogenetic modelling of the variation of the evolutionary process across sites from multi-species sequence alignments has garnered increasing attention over the last few decades. One of the main approaches, sometimes known as *random effects* modelling, adopts the view that the heterogeneity across observations is a result of the data set having been emitted from several different models, each drawn from a distribution. When little is known about the form of the across-site heterogeneity, *finite mixture* models provide discretizations of the unknown distribution into a pre-determined set of sub-models, or components. Choosing a level of discretization that is sufficiently fine-meshed to reflect the underlying heterogeneity is typically done from a set of likelihood-based model comparisons using different numbers of components. In the *infinite mixture* framework, accounting for the uncertainty regarding the number of components is another layer built into the model formulation (i.e., a hierarchical modelling framework), providing a rich non-parametric fitting of the distribution of across-site heterogeneity. Here, we use Bayesian cross-validation to compare a wide range of finite mixture models, along with the infinite mixture modelling approach known as categories, 'CAT', and gamma-distributed rates-across-sites approach. We study the model comparison approach on simulations, and apply it to five real multi-gene alignments. Our findings indicate that the potential improvement in model-fit from finite mixture models is attained when the number of components of the mixture is between 20 and 60. The magnitude of improvement from the mixture model is highly dependant on whether or not the gamma-distributed rates-across-sites approach is invoked. Moreover, in all cases that we considered, the fit of the CAT-GTR+Γ model matched or exceeded the best-fitting finite mixture model.

Table of Contents

List of Figures

List of Tables

1. Introduction

Fundamentals of Phylogenetics

The modern process of determining the relatedness of different groups of related organisms starts with collection of DNA from specimens of each species of interest. This collection can be the entire genome, specific genes present in all organisms under study or other collections of genetic information. Regardless of the genetic markers used, the DNA is collected and combined into a multiple sequence alignment. The multiple sequence alignment is then used to infer a probable topology which represents the evolutionary history of the specimens collected. A example of a possible tree topology given a multiple sequence alignment is given in figure 1.

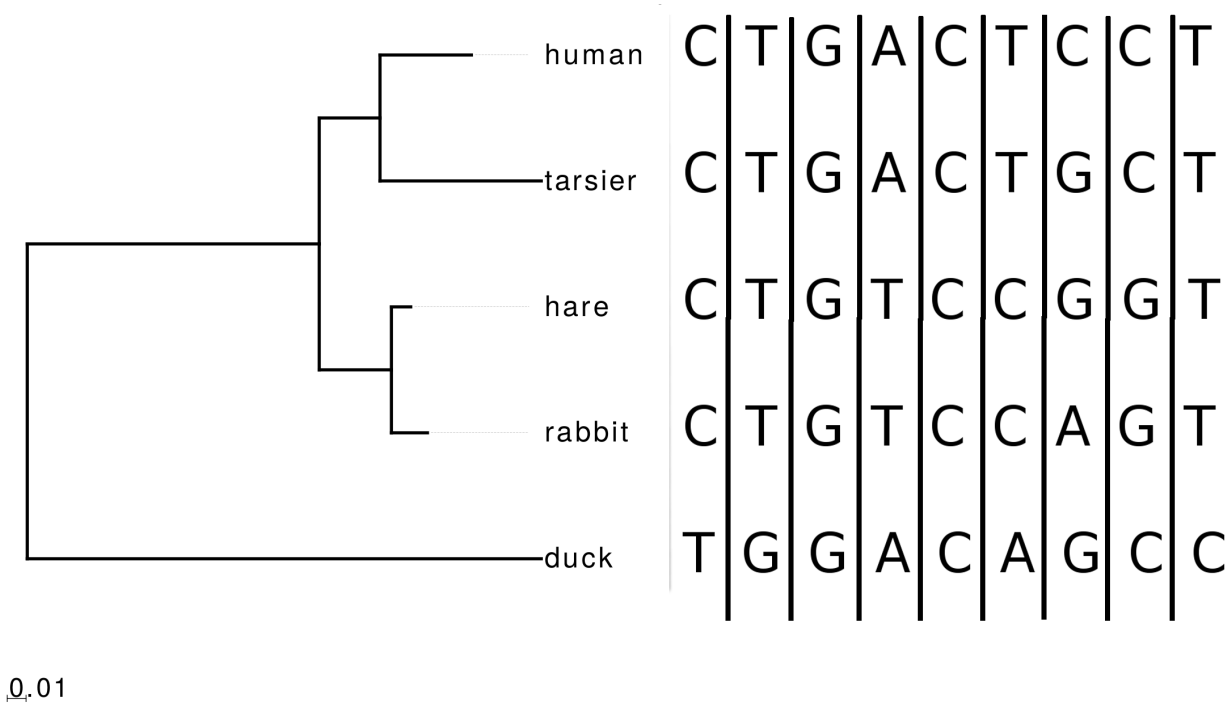

Figure 1. Example of a phylogenetic tree and nucleotide multiple sequence alignment

Multiple sequence alignments are DNA sequences collected into a matrix where each row represents a different species and each column (position) represents the nucleotide (or amino acid, or codon) state across multiple organisms. There are many methods which can be used to reconstruct the phylogeny of aligned genes. The most used are probabilistic phylogenetics methods including maximum likelihood phylogenetic inference (Felsenstein, 1981) and Bayesian phylogenetic inference (Li et al., 2000; Yang and Rannala, 1997; Larget and Simon, 1999).

Different models are used to perform phylogenetic analysis on multiple sequence alignments. These models vary in their complexity, but nearly have some common parameters. The substitution matrix, Q, is used to represent the infinitesimal rates at which one state (nucleotide or amino acid or codon) changes to another. The entries in this rate matrix are specified from two sets of parameters, and are given by:

$$Q_{ij} = \rho_{ij}\pi_j. \tag{1}$$

Here, ρ_{ij} is the exchangeability between the initial state, i, and the final state, j. Exchangeability parameters can be inferred from the data set under analysis, or can be taken from empirical studies conducted beforehand. The exchangeability value is multiplied by the frequency of the final state, represented by π_j.

A common phylogenetic model, and a primary model used in experiments in this paper, is the General Time Reversible (GTR) model (Tavaré, 1986). The GTR model has symmetrical exchangeabilities, meaning that the same value is used for an exchange from A to T as from T to A, or any other pair of states; in other words $\rho_{ij} = \rho_{ji}$. All of these symmetrical

exchangeability parameters are inferred from the data set. The frequency parameters, π_j, are also inferred from the data under study, either by maximum likelihood or Bayesian inference.

Empirical models are often used in amino acid level data because they have been shown to have adequate fit and can decrease run time due to less parameters having to be inferred. A typically used set of empirical exchangeability values are those of the Whelan and Goldman (2001) matrix, displayed in figure 2.

The simplest possible substitution model, referred to as Poisson in this paper, treats all exchanges as having equal probability (Jukes et al., 1969; Tavaré, 1986). It is also possible have intermediate approaches. For instance, the F81 model (Felsenstein, 1981) assigns equal exchangeability parameters between all pairs of states, but infers the frequency parameter values.

Likelihood Calculation

Probabilistic phylogenetics rests upon the likelihood function. This function is defined as the probability the data (D) at hand given all parameters of the models being used. In symbolic notation, we write $p(D|\theta)$, where θ collectively denotes all parameters. The first step of computing the likelihood is the calculation of the probability of transitioning from one state to another at a particular position over a evolutionary time period, or branch length (λ, representing the expected number of substitution per site along the branch). This probability can be calculated from the rate matrix, Q, by matrix exponentiation:

$$p(G|A,\theta) = \left[e^{\lambda Q}\right]_{AG}, \tag{2}$$

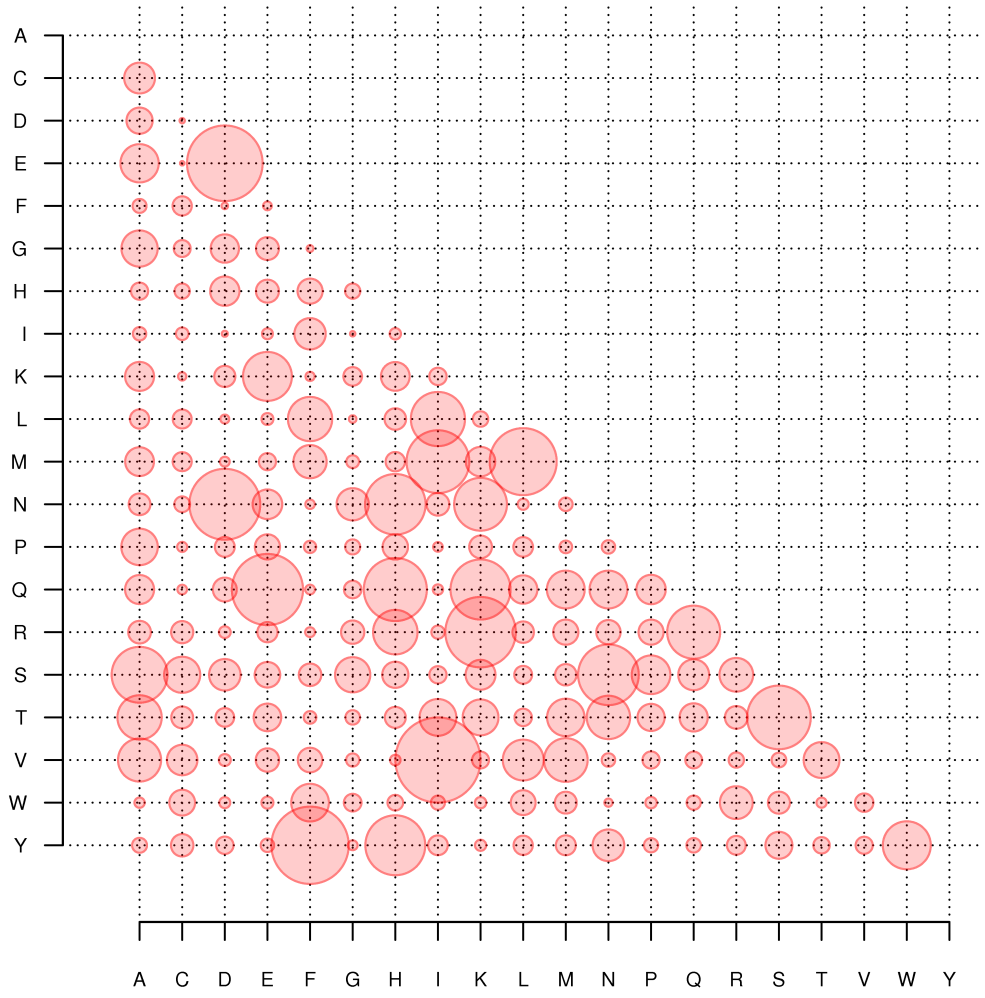

Figure 2. Relative rates of exchangeability from Whelan and Goldman (2001). Larger circles represent higher relative rates of exchangeability between two amino acids.

where A represents the initial state, G represents the final state. This step is central to the likelihood calculation.

The likelihood calculation requires that we compute the transition probability from one state to another for all branches of a particular internal node state configuration. An internal

node state configuration is an arbitrarily chosen set of ancestral states at each node in the phylogeny. The transition probabilities across all branches for a particular internal node state configuration are then multiplied together. An example of a possible internal node state configuration with transition probabilities is given in figure 3.

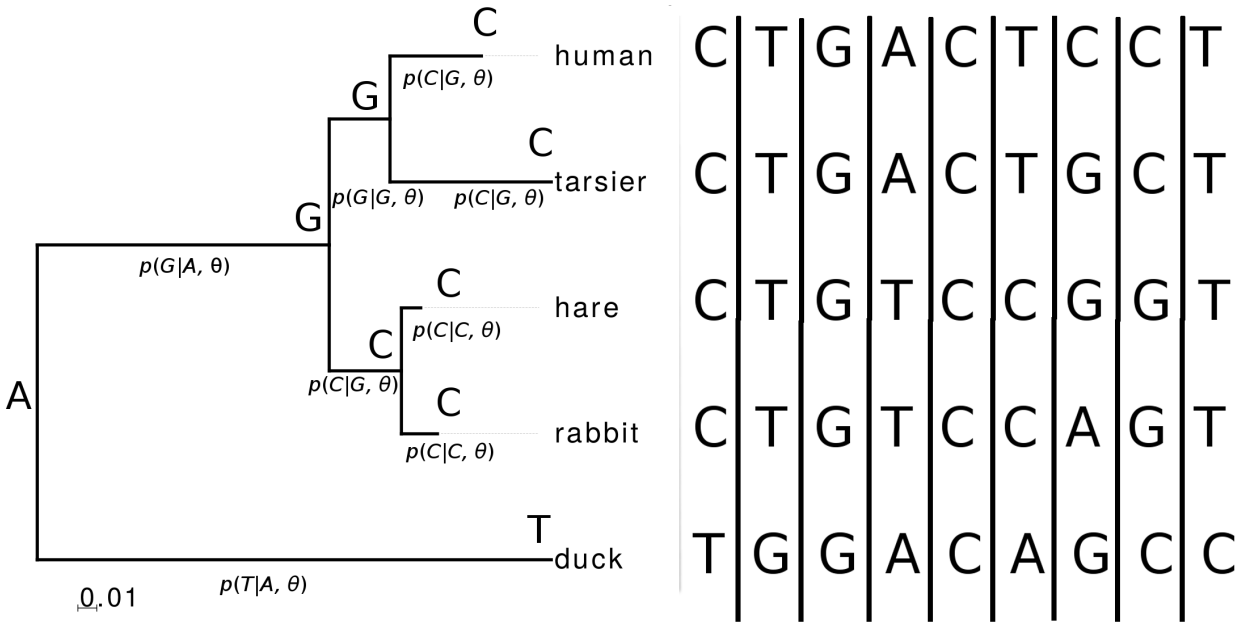

Figure 3. Phylogenetic tree with a possible internal node state configuration

The internal node states are not known. For this reason, the process above is repeated for every possible node state configuration. The products of transition probabilities for every internal node state configuration are then summed, giving the likelihood value for the site under consideration. As a final step, the processes described above are repeated for every site in an alignment and the product of these site likelihoods is the overall likelihood.

A summary of the likelihood calculation follows:

1. Use the matrix exponentiation step (equation 2) of the likelihood calculation to calcu-

late the probability of transition from a starting state at a node to the ending state;

2. Apply step 1 on all branches for a particular node state configuration;

3. Repeat step 2 for every internal node state configuration;

4. Sum the products for every internal node state configuration;

5. Repeat previous four previous steps for all sites and take product across sites.

There are some assumptions inherent to the likelihood calculation, most of which are motivated by reducing the computation time required. One assumption is that there is independence of evolution between sites. This assumption, though likely not realistic due to the possible interactions between amino acids encoded by the underlying DNA sequence, is what allows us to take the products of site likelihoods. Another assumption is that of independence across lineages, or branches. Again species are likely to have evolutionary interactions, but the assumption of Independence allows for taking of the product of transition probabilities.

In maximum likelihood phylogenetic inference, the parameters are adjusted in order to determine the set of values that results in the highest likelihood score (Felsenstein, 1981). In Bayesian phylogenetics the likelihood is combined with a prior probability in order to compute the posterior probability. In this paper, we focus on Bayesian phylogenetics.

Bayesian Phylogenetics

The primary difference between maximum likelihood phylogenetics and Bayesian phylogenetics is that maximum likelihood only gives a point estimate on parameters. Bayesian

phylogenetics provides a distribution of estimates for a parameter. This distribution of estimates is conditional on the data. There are benefits and drawbacks to each method. Maximum likelihood methods can be significantly faster than Bayesian methods, which allows for phylogenetics with vast quantities of data such as in whole genome phylogeny. Bayesian phylogenetic analysis is the method primarily used for studying more complex models of substitution, such as those with across-site heterogeneity.

The basics of Bayesian statistics is the equation of conditional probability (Bayes' theorem):

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} \tag{3}$$

There are four expressions in Bayes' theorem. First, $p(\theta)$ is the prior probability. This prior probability is meant to represent the investigator's degree of belief in parameter values before having considered any data. In practice, prior probabilities are often defined by the developers of the Bayesian software to be used. With sufficient data, prior definitions do not change results, though with small data sets the effect of the prior on results must be carefully considered. The second expression, $p(D|\theta)$, is the likelihood function, which we have discussed how to calculated previously. The third expression, $p(D)$, is the marginal likelihood, which serves to ensure that the total posterior probability equals 1. Finally the posterior probability, $p(\theta|D)$, represents our degree of belief in the parameter values after considering the data at hand.

The marginal likelihood cannot be calculated analytically, which means that the posterior probability cannot be calculated analytically. Though it is not possible to calculate the

posterior, it is possible to sample from it. The most common method of sampling from the posterior is the Markov chain Monte Carlo (MCMC) algorithm.

The general idea behind MCMC is to construct a 'random walk' in the space of all possible sets parameter values. By this we mean that given a particular set of parameter value, θ, we move to another set of parameter values, θ'. Such moves are repeated multiple times over the course of the random walk. This random walk is biased according to the Metropolis-Hastings rule (Metropolis et al., 1953; Hastings, 1970). The bias is constructed so as to visit sets of parameter values that have probability more frequently than sets of parameter values that have low probability. The MCMC algorithm is as follows:

1. Draw an initial set parameter values, θ, from the prior;

2. Change current parameter values (mechanisms for such changes are reviewed in Rodrigue and Lartillot, 2012), these new parameters are θ';

3. Compute the acceptance probability, $\vartheta = min\{1, \text{MH-ratio}\}$, where MH-ratio is the Metropolis-Hastings ratio expression, equal to $\frac{p(\theta'|D,M)}{p(\theta|D,M)} \frac{q(\theta',\theta)}{q(\theta,\theta')}$; f

4. If $\vartheta = 1$, replace θ with θ'; If not, draw a random value from a uniformly distributed unit interval (i.e., between 0 and 1) and replace θ with θ' if this draw is less than ϑ;

5. Record θ;

6. Go to step 2.

By looping over these steps a very large number of times, we gather a large sample of sets of parameter values. High posterior probability values occur more frequently in our

sample. In fact, the frequency of occurrence of a set of parameter values in our sample is our approximation of the posterior probability of that set of values.

It is important to note that the key to the algorithm relies on the Metropolis-Hastings ratio in which the marginal likelihood cancels:

$$MH = \frac{\frac{p(D|\theta')p(\theta')}{p(D)}}{\frac{p(D|\theta)p(\theta)}{p(D)}} \frac{q(\theta', \theta)}{q(\theta, \theta')} \tag{4}$$

$$= \frac{p(D|\theta')p(\theta')}{p(D|\theta)p(\theta)} \frac{q(\theta', \theta)}{q(\theta, \theta')}. \tag{5}$$

The Hastings ratio, $\frac{q(\theta',\theta)}{q(\theta,\theta')}$, serves to correct for biases in the proposal mechanism from step 2 (see Rodrigue and Lartillot, 2012). The long term behaviour of the algorithm is independent of the proposal mechanisms used and the initial draw from the prior. In practice, given our finite sample size from the posterior, the first cycles (also known as the 'burn-in') from the algorithm are generally removed due to them not being in the equilibrium state of the MCMC.

Across-Site Variation and Phylogenetic Modelling

Gamma Distributed Rates Across Sites

When working with phylogenetic data, one of the implicit assumptions that is made in simple models is that all of the sites from an alignment are governed by the same set of parameters. This is unlikely to reflect the underlying processes in biological systems. There are many cases where different base pairs in DNA sequences have vastly different selective pressures acting on them. For example, some sections of DNA may code for hydrophobic sections of proteins, therefore these sections will not have high chances of transitioning into

codons which encode for polar amino acids (Echave et al., 2016). There are many examples of this selectivity, such as selection for a certain charge on amino acids or selection for amino acid size, aromaticity, sulfide presence, and particular amino acids key to protein shape. Attempting to account for the complex evolutionary patterns across sites was the impetus for the creation of the gamma-distributed rates-across-sites model (Yang, 1993).

The gamma-distributed rates-across-sites is used in phylogenetics to allow for heterogeneity of overall rates to be introduced into a model. It allows different sites to have different branch length multipliers, which, in effect increase or decrease branch lengths for those sites. This approach does not distinguish between the states of the substitution process, but simply increases or decreases the rates of substitution across sites. The gamma-distributed rates-across-sites model is applied by treating these branch length multipliers as random variables from a gamma distribution of mean 1 and of variance $1/\alpha$ [1]. The parameter α, which controls the shape of the distribution, becomes part of the overall inference. The distribution is a discrete approximation of the gamma distribution. Under this model, the likelihood function at a particular site i becomes a weighted average of the likelihood across all possible rate values at that site, r_i, permissible by the gamma law;

$$p(D_i|\theta) = \int_{r_i} p(D_i|r_i, \theta) p_\alpha(r_i) dr_i. \tag{6}$$

In practice, the integral given in equation 6 has no analytical solution and is therefore approximated through a discretization approach. This reduces the integral into a weighted

[1]The gamma distribution has two parameters which determine its properties, α and β. The shape parameter is α and β is the rate parameter. The gamma distribution has mean of α/β and variance α/β^2. The mean of the gamma distribution is restricted to 1 by holding β to be equal to α. This simplifies the variance to $1/\alpha$.

sum:

$$p(D_i|\theta) = \sum_{k=1}^{K} p(D_i|r_k, \theta)w_k, \tag{7}$$

where w_k is a weight associated to the kth class of the discretization. Typically the discretization is done so as to have four classes which is a good compromise between computational costs and a suitable rendering of the gamma distribution (Yang, 1994). When invoked, this approach is denoted with the suffix $+\Gamma$, as in GTR+Γ.

Partitioning

Partitioning is another method of allowing variation across sites in probabilistic phylogenetics. Partitioning involves explicit *a priori* definitions of the the sites of an alignment which are believed to have the same underlying evolutionary pressures in a collection known as a data block (Kainer and Lanfear, 2015). When working with protein coding DNA, these partitions typically fall along the sites which are from introns or exons, different genes in the alignment and different codon positions in those genes (Shapiro et al., 2005). For amino acid level data, partitioning is generally limited to data blocks containing entire proteins. Each data block uses a separate model from the others for inference and the parameters as part of the inference.

Partitioning can be a valuable method of introducing heterogeneity to probabilistic phylogenetic modelling but there are flaws which limit its utility. Partitioning, by relying on *a priori* site assignment, is limited in the data blocks which can reasonably be created. For instance, with amino acid data, partitioning completely overlooks heterogeneity within proteins.

Mixture Models

Another perspective on the issue of capturing across-site heterogeneity is to explicitly *model* the uncertainty in partitioning, by adopting a *random-effects* framework (Rodrigue and Lartillot, 2012). In this framework, the model considered to have generated a particular datum is itself drawn from a statistical law. For some random-effects-based modelling objectives, it is feasible to work with a characterized statistical law, such as in the gamma-distributed rates-across-sites models (Yang, 1993, 1994). For other objectives, when no obvious statistical law is forthcoming, finite mixture models provide a means of discretizing the unknown distribution into a predetermined number of components, each consisting of a set of amino acid frequencies, with respective weights (Pagel and Meade, 2004; Le et al., 2008b; Wang et al., 2008; Susko et al., 2018).

Finite mixture models were first explored on nucleotide-level data (Pagel and Meade, 2004), but have since been of great interest for models operating directly in the amino acid state space (Le et al., 2008b; Wang et al., 2008). The heterogeneity within a protein is highly pronounced in amino acid alignments, with numerous columns displaying signatures of substitution histories over a limited sub-set of amino acid states (see, e.g., Echave et al., 2016). Le et al. (2008b) worked with finite mixtures of amino acid profiles of up to 60 components, while assuming even amino acid exchangeability parameters. In effect, such a model has 60 different substitution matrices, as defined in equation 2, with each matrix having a distinct set of frequencies (π). Wang et al. (2008) worked with finite mixtures of profiles combined with free amino acid exchangeability parameters, but with relatively few components (4 or 5). Recent work by Susko et al. (2018) combined empirical amino acid

exchangeability parameters (Le et al., 2008b; Le and Gascuel, 2008) with finite mixtures on amino acid profiles in a maximum likelihood context, and suggest that generic finite mixtures of profiles (Le et al., 2008a) are surpassed by data-adjusted finite mixtures.

Infinite mixture models, such as those based on the Dirichlet process (Lartillot and Philippe, 2004), are a form of random-effects approach that is not restricted to the manifold of any particular parametric family of distributions. As such, it is often described as a non-parametric means of capturing across-site heterogeneity (Lartillot et al., 2007), although this should not be considered to imply that no parameters are involved in its specification. Rather, parameters are invoked to specify a prior distribution over a family of discrete distributions, allowing the model to flexibility 'pixilize' the true underlying distribution. Although conceptually more elaborate than finite mixture models, the infinite mixture models utilizing the Dirichlet process prior have been more extensively studied (e.g., Lartillot et al., 2007; Feuda et al., 2017).

Cross-Validation Model Comparison

There are many methods of comparing statistical models. A common method, such as the Akaike information criterion (AIC) (Akaike, 1974), can work equally as well as cross-validation in maximum likelihood context (Stone, 1977), though these methods are not as useful in a Bayesian context. These methods for model comparison do not explicitly test model predictive power, they instead penalize for introducing new parameters into the model. Cross-validation on the other hand explicitly tests for predictive power thus limiting model dimensionality inherently.

As discussed previously, there are many different types of substitution models. When performing an analysis on real data one requires a means of choosing the model which is most appropriate. Many different statistical approaches exist for this task (see Sullivan and Joyce, 2005). Cross-validation is a useful and general method for comparing the statistical merit of different models.

Cross-validation analysis begins with some of the data being set aside. This data (the test set) is not used until the end of the process. The remaining data (the learning set) is used to infer model parameters. Having obtained an inference of model parameters from the learning set, the parameter values are then used on the test data set. By calculating the likelihood function on the test data we get a cross-validation score for a specific model. In the Bayesian context the cross-validation score is averaged over all parameter values of our sample from the posterior distribution:

$$p(D_2|D_1) = \int_\theta p(D_2|\theta)p(\theta|D_1)d\theta, \tag{8}$$

where D_1 is the learning data set, D_2 is the testing data set and θ is the set of parameters for a specific model. The posterior in equation 8, $p(\theta|D_1)$, is incalculable though it can be sampled via the MCMC approach resulting in a collection of parameter values, $\theta^{(k)}, 1 \leq k \leq K$. This sample of K sets of parameter values can then be used to approximate equation 8 as a summation:

$$p(D_2|D_1) \approx \frac{1}{K}\sum_k p(D_2|\theta^{(k)}). \tag{9}$$

where $p(D_2|\theta^{(k)})$ is the likelihood on the test data.

Cross-validation schemes can split the data in many different way. Common methods to split the data are 1/2 (half of the data as the test set and half as the learning set) and 1/10th (one tenth of the data as a test set, nine tenths of the data as the learning set). The process of separating the data into testing and learning data sets is usually done at random, and repeated multiple times. In such cases, the mean cross-validation score and the standard deviation are reported.

Here, we use Bayesian cross-validation to produce a ranking of both finite and infinite mixture models accounting for across-site variation in amino acid frequency parameters (or *profiles*, for short). We expect to find that infinite mixture models provide a improved fit over homogeneous models, however the level of complexity of finite mixture models required to compete with infinite mixture models is difficult to foresee. We also attempt to evaluate the relative merit of two different modelling strategies: gamma-distributed rates-across-sites (Yang, 1993, 1994) and mixtures of profiles across sites (Lartillot and Philippe, 2004). We expect both strategies to yield good model fit, especially when these strategies are combined, however the relative improvements are difficult to estimate.

Our results confirm recent findings by Susko et al. (2018), in that finite mixture models always out-perform their homogeneous counterparts, and that data-adjusted finite mixture always out-perform the generic empirical finite mixtures of Le et al. (2008b). Improvement in model fit, provided by finite mixture models, is highly sensitive to Whether or not the gamma distributed rates model is invoked. We also find that infinite mixture models always match or out-perform the best-performing finite mixture models.

2. Materials and Methods

Data Sets

We selected 3 data sets from Kainer and Lanfear (2015) to work with. Selection was based on the size of the alignments in both number of taxa and number of sites. Computational resources restricted the total number of possible data sets which could be analyzed to 3. We make use of a shorthand to refer to our data sets, indicating the name of author, number of taxa, and number of sites

- Broughton-61-19997 - Broughton et al. (2013) - A concatenation of 20 nuclear genes and 1 mitochondrial gene from 61 different species of fish.
- Lartillot-78-15117 - Lartillot and Delsuc (2012) - 17 protein-coding genes aligned from 73 placental mammals.
- Wainwright-188-8439 - Wainwright et al. (2012) - DNA sequences collected from 10 protein coding nuclear genes from 188 species of perch-like fishes.

Sequences were converted into amino acid format from their original nucleotide format. These data sets are sufficiently large thus the prior is not expected to effect the final results of cross-validation analysis.

Substitution Models

We used 5 types of substitution models. The first is the general time-reversible models (GTR) (Tavaré, 1986), which is a homogeneous substitution process across sites. The second is the CAT model, which is an infinite mixture of state profiles across sites, but with

flat (Poisson) exchangeabilities between states. The CAT-GTR model combines the homogeneous substitution processes across sites and infinite mixtures (Lartillot and Philippe, 2004). We also use finite mixture models, which we refer to as fixed component CAT-GTR (CAT_f-GTR) which are models where we can specify the exact number of components with profiles and weights estimated, and where we explore several values for the number of components. Finally we used empirical profile mixture models (C20, C40 and C60) (Le et al., 2008b). In addition to this, we invoked the gamma-distributed rates approach, with 4 discrete categories (Yang, 1994). We also analyzed models which suppressed the gamma-distributed rates-across-sites.

Cross-validation

We used a 5-fold, 5-replicate cross-validation approach to compare substitution models. This procedure randomly splits data sets into two parts, the learning (or training) part and the testing part. The training data set, under five fold cross-validation, contains four-fifths of sites from the original data set (chosen at random from the whole data set) and the testing set contains the remaining one-fifth. The model parameters, inferred from the learning set are then used to analyze the test data set, a data set which, in effect, the model has never seen before. In this way, the test provides a genuine measure of the predictive power of a given model. This process is computationally intensive, but allows for comparison of any set of models of interest. Five fold cross-validation was chosen to ensure that the learning data set was significantly larger than the testing data set while still having a significant portion of the data available for testing.

Cross-validation analysis on the real data sets was carried out with training data that was run until 1000 cycles of the MCMC occurred in the Phylobayes software (Lartillot et al., 2009). The values of the likelihood were visually inspected for convergence. The most complex model used in this analysis, CAT-GTR+Γ, was used to assess the required number of cycles required for suitable convergence. Tree topology was also treated as a free parameter rather than specified, a decision which does increase computation time. Running cross-validation under a fixed topology reduces computation time and should not cause less significant results as long as the models compared are sufficiently distinct. The cross-validation scores were computed without the 400 burn-in cycles. In effect the analyses used 600 MCMC cycles to compute cross-validation score.

Cross-validation analysis on the simulated data sets followed the same procedure as that of the real data analysis with the exception of the number of cycles. The simulations used 500 cycles with a burn-in of 400, leaving 100 effective cycles for analysis to be carried out on. The number of cycles used after burn-in for both real data and simulation analyses were selected due to computational time constraints. The limited number of cycles used did not hinder the robustness of the cross-validation analysis results as multiple independent replicates yielded similar results.

Following analysis, cross-validation scores were calculated and compared relative to GTR+Γ cross-validation score for all models analyzed.

Simulations

To test how these models performed when the evolutionary signal was increased or decreased, we generated data using the tree topology obtained from the Lartillot-78-15117. This tree was generated using phyML (Guindon and Gascuel, 2003) with the GTR+Γ model. Some simulations used this topology and branch length directly, whereas others had branch lengths multiplied by a factor of 10 or 0.1. The variable branch lengths were used to assess how the models performed on substitution rich and substitution poor data sets. This information can be used to provide a basis for interpreting results from real data. By observing the cross-validation results under different levels of substitution we can determine which cases allow for accurate model comparison. Following the multiplication of branch length values, alignments were generated using Whelan and Goldman (2001) exchangeabilities and different empirical frequency sets from Le et al. (2008a). The frequency sets used were C20, C40, and C60. Each alignment was 6000 amino acids long. When simulating with the C20 profiles, 300 positions were produced using each of the 20 sets of profiles. When simulating with 40 profiles, 150 positions were produced using each of the 40 set of profiles. Finally, When simulating with 60 profiles, 100 positions were produced using each of the 60 set of profiles. composed of 20, 40 or 60 different alignments generated under each of the C20, 40 or 60 frequency data sets. The gamma rate heterogeneity was not invoked in these simulations.

For example, the parameters used to generate panel A in figure 4, was the Latilloit-78-15117 topology generated from a phyML Guindon and Gascuel (2003) analysis. This tree had its branch lengths multiplied by an order of magnitude down. This tree was used as the base tree in Seq-gen. Each of the 20 empirical profiles was used with length of 6000/20

given to that frequency set. The parameters used to generate each portion of the total 6000 amino acid alignment were; the WAG general rate matrix, a continuous gamma rate parameter and one of the frequency parameters from Le et al. (2008a). After generation of all 20 alignments, with the different empirical frequencies from C20, they were concatenated into one single alignment 6000 amino acids in length. This alignment was the one used for cross validation analysis under the models GTR+Γ, CAT-GTR+Γ and the CAT$_f$-GTR+Γ.

Cross validation analysis under these simulations was carried out on chains that were ran for 500 cycles instead of the 1000 used for the read data sets. This change was made due to time constraints. A burn-in of 400 cycles was still used, making the effective cycles used for analysis 100.

3. Results and Discussion

Simulations

We first explored the Bayesian cross-validation framework using simulations. Our objective here was to assess if the Bayesian cross-validation can recover the appropriate finite mixture model when applying models treating amino acid exchangeabilities and all aspects of the mixture as free parameters: if a sufficiently rich data set was produced with 20 components, for instance, a finite mixture model with 20 components should be elected as most appropriate.

Figure 4 shows example results of several model comparisons on simulated data sets, with the abscissae corresponding to the number of components in finite mixture models, and the ordinates are the Bayesian cross-validation scores relative to the GTR+Γ model. The top panels (A, B, and C) report results on simulations with C20-WAG, the middle panels (D, E, and F) on C40-WAG simulations, and the bottom panels (G, H, and I) on C60-WAG simulations; left panels were based on simulations with branch lengths one-tenth of the original tree branch length, middle panels with the original branch lengths, and right panels with branch lengths ten times those of the original tree. At very shallow evolutionary depths, such as with the simulations conducted over the tree with branch lengths one-tenth of the original branch length, invoking the finite mixture models provides a very weakly improved cross-validation score over the GTR+Γ model (panels A, D, and G), with their difference in score only slightly above 0 (and error bars often encompassing 0). The C60-WAG simulations are the only case, with one-tenth branch lengths giving a statistically significant improvement (panel G). This result is explained by the fact that over such short

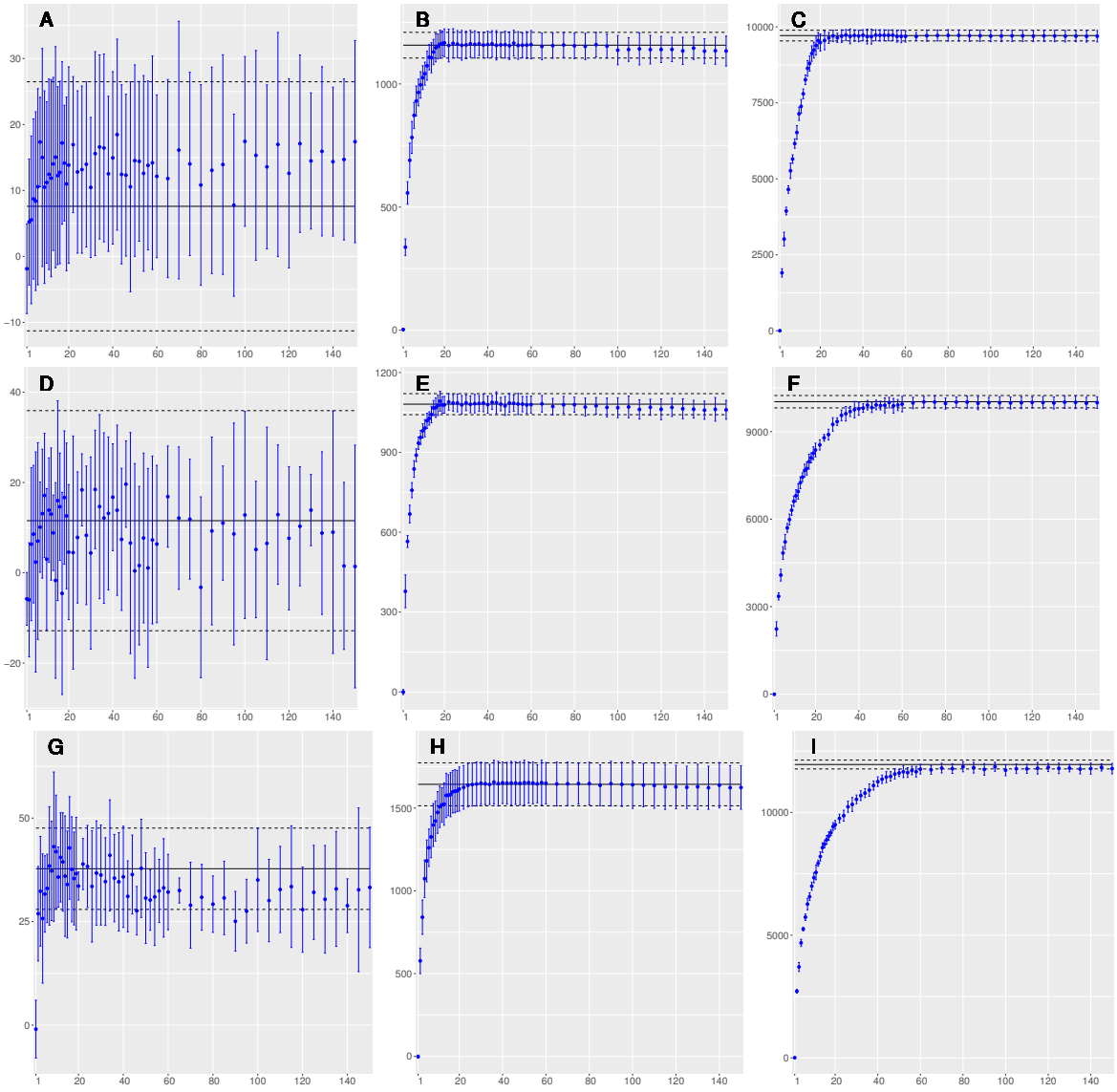

Figure 4. Cross-validation scores for simulated data. Solid black lines represent scores for CAT-GTR+Γ models, dotted black lines represent the associated standard deviation. Blue points represent score for the CAT$_f$-GTR+Γ models with component numbers (f) corresponding to their associated x-axis position, blue error bars represent standard deviation for each CAT$_f$-GTR+Γ model. Each panel represents a particular model and set of branch lengths as simulation conditions: C20-WAG at one-tenth branch lengths (panel A); C20-WAG with original branch lengths (panel B); C20-WAG at ten times branch lengths (panel C); C40-WAG at one-tenth branch lengths (panel D); C40-WAG with original branch lengths (panel E); C40-WAG at ten times branch lengths (panel F); C60-WAG at one-tenth branch lengths (panel G); C60-WAG with original branch lengths (panel H); C60-WAG at ten times branch lengths (panel I).

evolutionary distances, the substitution process defined by the C20-WAG, C40-WAG, and C60-WAG models has not been actualized; with so few substitutions simulated, the Bayesian cross-validation score suggests the use of the comparatively compact GTR+Γ model. Even with the simulations conducted with un-altered branch lengths (panels B, E, H), the Bayesian cross-validation scores indicate that a model with fewer components than the true generative model used to simulate is preferred. Nonetheless, there is a clear preference for mixtures models over the plain GTR+Γ model. It is only with the simulations conducted over a tree with branch lengths multiplied by 10 that the best-performing finite mixture model matches the true generative model in number of components, with the Bayesian cross-validation score reaching a plateau between 20 and 25 components for the C20-WAG simulations (panel C), between 40 and 45 for the C40-WAG simulations (panel F), and between 60 and 65 with C60-WAG simulations (panel I).

The plateau of cross-validation scores reached is itself noteworthy. Although Bayesian cross-validation implicitly penalizes for model dimensionality, it seems this natural penalty is often weak; once a sufficiently rich mixture model is invoked, adding more components does not provide any improvement. Finite mixture models, though the Bayesian framework, adjust the parameter values by assigning a weight to them near 0, so as to suppress superfluous components. In general, over-parameterization is less of a issue in Bayesian contexts (Efron, 2005). Also noteworthy is the fact that the plateau of cross-validation scores reached with finite mixture models is at CAT-GTR+Γ levels. The flexibility of the Dirichlet process apparatus circumvents any time-consuming model comparisons of finite mixtures, always leading to the top-scoring configuration in a single run.

Altogether, these experiments indicate that the Bayesian cross-validation procedure per-

forms well, but in order for rich substitution models to be fully expressed in simulations, a great deal of evolutionary signal is required. With information-poor data sets, finite mixtures with fewer components, or even homogeneous substitution processes across sites, become preferred. Finally, the CAT-GTR+Γ model automatically adjusts to a configuration matching the best-fitting finite mixtures, or essentially reverting to the plain GTR+Γ substitution model if no particular heterogeneity is warranted.

Real Data Analyses

We next conducted the Bayesian cross-validation procedure on real data sets. Figure 5 is a compilation of all the data gathered through cross-validation analyses on the Broughton-61-19997 data set. We summarize the models considered below:

- GTR: Model with free exchangeabilities and no gamma-distributed rates-across-sites;
- GTR+Γ: Model with free exchangeabilities and gamma-distributed rates-across-sites;
- CAT-Poisson: Infinite mixture model with equal (Poisson) exchangeabilities and no gamma-distributed rates;
- CAT-Poisson+Γ: Infinite mixture model with equal (Poisson) exchangeabilities and gamma-distributed rates;
- CAT-GTR: Infinite mixture model with free exchangeabilities and no gamma-distributed rates;
- CAT-GTR+Γ: Infinite mixture model with free exchangeabilities and gamma-distributed rates;

- CAT$_f$-GTR: Finite mixture model with f components, free exchangeabilities and no gamma-distributed rates;
- CAT$_f$-GTR+Γ: Finite mixture model with f components, free exchangeabilities and gamma-distributed rates;
- C20-GTR: C40-GTR: C60-GTR: Empirical mixture models with free exchangeabilities and no gamma-distributed rates;
- C20-GTR+Γ: C40-GTR+Γ, C60-GTR+Γ, Empirical mixture models with free exchangeabilities and gamma-distributed rates;
- C20-Poisson, C40-Poisson, C60-Poisson, Empirical mixture models with equal (Poisson) exchangeabilities and no gamma-disturbed rates;
- C20-Poisson+Γ, C40-Poisson+Γ, C60-Poisson+Γ, Empirical mixture models with equal (Poisson) exchangeabilities and gamma-distributed rates.

GTR and GTR+Γ

In cross-validation analysis, different models are compared to a specific reference model. The reference model used for all cross-validation analyses was the GTR+Γ model. This model was chosen as the reference due to its prevalence in phylogenetics. All cross-validation scores reported are in fact the difference between the natural log-likelihood cross-validation score of a particular model and the cross-validation score of GTR+Γ.

As can be seen in figure 5, the GTR model, without any accounting of heterogeneity across sites, gives a much poorer score than GTR+Γ, by over 2000 natural logarithmic

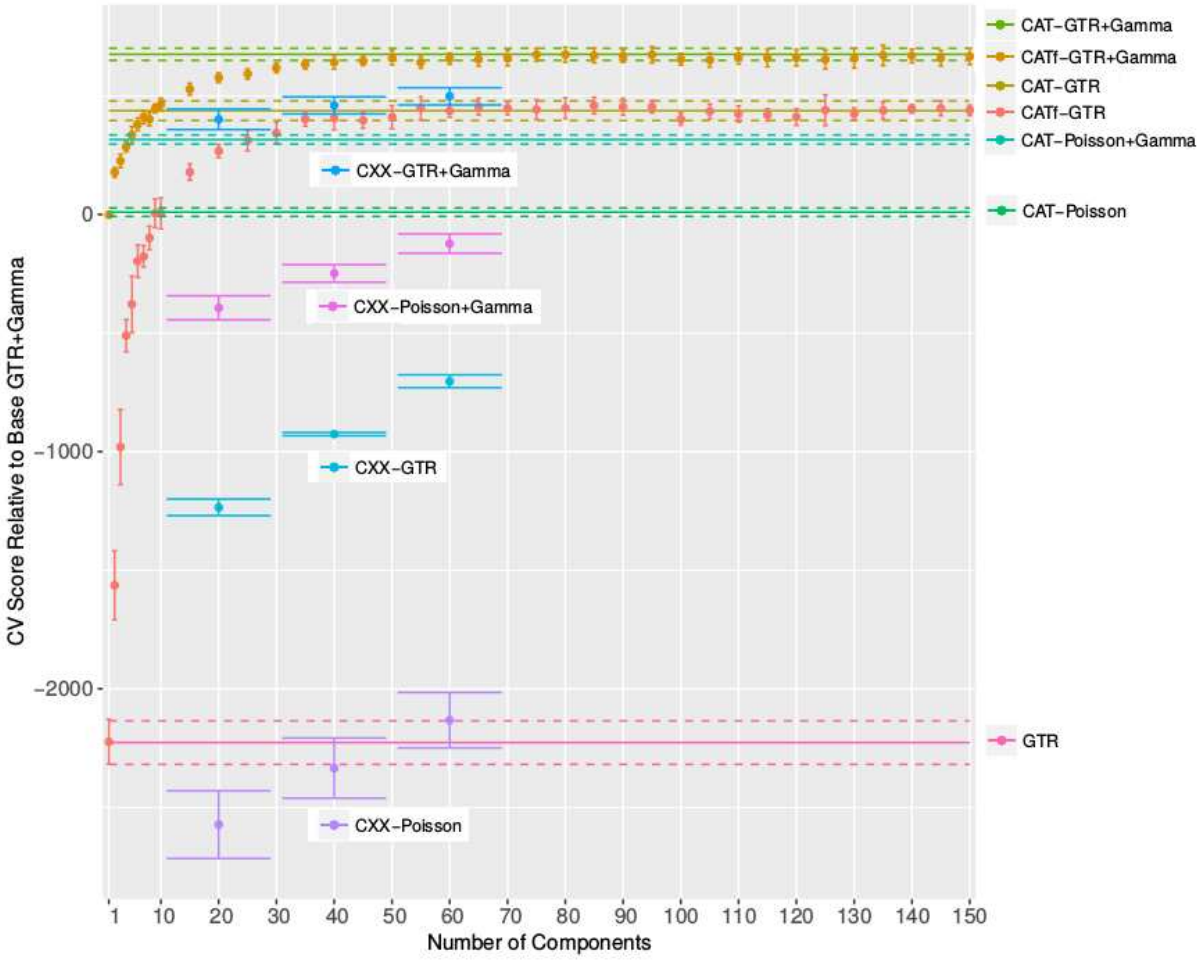

Figure 5. Cross-validation scores on the Broughton data set for all analyzed models

units. It has long been known that inclusion of gamma-distributed rates-across-sites into phylogenetic models improves the fit (Yang, 1996). This holds true in our experiments.

CAT-Poisson and CAT-Poisson+Γ

The CAT-Poisson model cross-validation score was only slightly above zero, with the standard deviation across the five replicates almost encompassing zero. This shows that the CAT-Poisson model, without Γ, only has slightly better model fit than GTR+Γ, though the improvement does not appear to be statistically significant. It is interesting that through two

completely different modelling rationales, one which accounts for heterogeneity in amino acid profiles (CAT-Poisson), and another which models variation in rates-across-sites (GTR+Γ), the overall model fit is nearly equivalently. In the other data sets analyzed (see table 1), CAT-Poisson outperforms GTR+Γ for one of the data sets (Wainwright-188-8439), and performs worse for the other (Lartillot-78-15117). The CAT-Poisson+Γ had notably higher cross-validation score than GTR+Γ and CAT-Poisson. The magnitude of the difference between CAT-Poisson and CAT-Poisson+Γ (306.18) was much smaller than the difference between GTR and GTR+Γ (2226.74). The smaller difference highlights how even though the CAT-Poisson model makes some simplifications in its assumption that all of exchangeabilities are equal, the treatment of the data as a heterogeneous collection of sites with their own substitution matrices has a large effect on model fit. CAT-Poisson, by treating the data as a heterogeneous collection of sites, which get assigned to specific categories, appears to occupy some of the benefit that is gained by adding in gamma-distributed rates-across-sites. In other words, invoking the gamma-distribution of rates-across-sites has a much weaker impact when working with mixture models, than when working with the GTR model. This effect is seen when comparing the difference in cross-validation score between CAT-Poisson and CAT-Poisson+Γ and how it is much smaller than the difference between GTR and GTR+Γ. The addition of CAT to a model is known to result in improved model fit (Lartillot and Philippe, 2004, 2006; Lartillot et al., 2007) and these analyses of CAT-Poisson and CAT-GTR are consistent with literature in showing that use of CAT in a model is an improvement over the homogeneous version of that model.

CAT-GTR and CAT-GTR+Γ

Final among the infinite mixture models, we have CAT-GTR and CAT-GTR+Γ. For every cross-validation analysis performed CAT-GTR+Γ was the best performing model, or matched any other model performance (see table 1). The difference in score between CAT-GTR and CAT-GTR+Γ was much smaller than the difference between between GTR and GTR+Γ for all data sets analyzed (see table 2). Specifically for the Broughton-61-19997 data set used to make figure 5, the difference between CAT-GTR and CAT-GTR+Γ was one-tenth the difference between GTR and GTR+Γ. When comparing the different infinite mixture models, it is apparent that CAT-GTR outperformed CAT+Γ. This indicates that, at least for infinite mixture models, the use of free exchangeability parameters is of greater importance than including a gamma-distributed rates approach. Although, as previously mentioned, combining free exchangeabilities and rate heterogeneity into with the infinite mixture model provides the highest fit. This result holds for all data sets (table 1).

Table 1. Cross-Validation Score For Every Model and Data Set Using GTR+Γ as Reference

	Broughton	Lartillot	Wainwright
GTR	-2226.7±91.9	-2013.7±81.4	-1342.7±168.7
GTR+Γ	Reference	Reference	Reference
CAT-Poisson	9.7±7.6	-245.9±112.0	167.7±53.6
CAT-Poisson+Γ	315.9±19.3	-134.0±125.3	353.5±80.5
CAT-GTR	437.4±41.4	471.5±54.1	404.9±92.7
CAT-GTR+Γ	675.3±25.5	554.0±39.5	1645.9±479.9
CAT_f-GTR	459.1±35.2	478.1±58.0	398.8±52.9
CAT_f-GTR+Γ	673.5±31.8	554.4±47.7	469.5±58.1
C20-GTR	-1234.9±34.9	-1243.9±50.4	-699.3±131.1
C40-GTR	-926.4±6.8	-1083.5±41.4	-580.8±135.1
C60-GTR	-703.3±27.1	-985.2±56.7	-520.2±120.8
C20-GTR+Γ	401.2±43.3	258.7±20.8	337.4±56.9
C40-GTR+Γ	460.1±36.1	343.5±20.9	376.8±66.3
C60-GTR+Γ	497.9±36.3	344.6±26.7	398.4±68.6
C20-Poisson	-2573.1±142.6	-3263.4±102.6	-1478.1±138.1
C40-Poisson	-2334.5±126.7	-3014.6±100.9	-1334.8±134.9
C60-Poisson	-2132.6±117.0	-2822.7±110.2	-1246.4±134.9
C20-Poisson+Γ	-393.4±50.7	-1363.5±64.2	-67.7±55.0
C40-Poisson+Γ	-248.9±37.1	-1196.7±55.7	39.0±61.6
C60-Poisson+Γ	-122.9±40.9	-1058.6±77.9	105.3±51.1

Note: CAT_f-GTR and CAT_f-GTR+Γ correspond to the finite mixture models with the number of components where cross-validation scores were maximized

Table 2. Relative cross-validation score improvement gained by including Γ

	Broughton	Lartillot	Wainwright
GTR vs GTR+Γ	2226.7	2013.7	1342.7
CAT-Poisson vs CAT-Poisson+Γ	306.2	111.9	185.7
CAT-GTR vs CAT-GTR+Γ	237.9	82.5	1241.0
CAT_f-GTR vs CAT_f-GTR+Γ	214.4	76.3	70.7
C20-GTR vs C20-GTR+Γ	1636.1	1502.6	1036.7
C40-GTR vs C40-GTR+Γ	1386.5	1427.0	957.7
C60-GTR vs C60-GTR+Γ	1201.2	1329.8	918.6
C20-Poisson vs C20-Poisson+Γ	2179.6	1900.1	1410.4
C40-Poisson vs C40-Poisson+Γ	2085.62	1818.0	1373.8
C60-Poisson vs C60-Poisson+Γ	2009.71	1764.0	1351.7

Note: Difference between CAT_f-GTR and CAT_f-GTR+Γ correspond to f values that had highest cross-validation score

CAT_f-GTR and CAT_f-GTR+Γ

We explored finite mixture models which treat all aspects of the mixture as free parameter. When using a finite mixture model, one must choose the number of components. Here, we explore many different values for the number of components (f), ranging between 1 (a non-mixture models) and 150 (a very rich mixture model), and compute the cross-validation score for each.

The cross-validation scores for the finite mixture models CAT_f-GTR and CAT_f-GTR+Γ follow a similar pattern across the range of f values. Both settings had cross-validation scores

which began at low values and rose quickly as the number of components was increased. As the number of components further increased the model began to encounter diminishing returns. For the Broughton-61-19997 data set (figure 5), these diminishing returns began occurring at roughly 10 components. The scores of CAT$_f$-GTR and CAT$_f$-GTR+Γ then plateau at similar values to those of CAT-GTR or CAT-GTR+Γ, respectively, at around $f = 40$ components. For the other two data sets the number of components required for CAT$_f$-GTR and CAT$_f$-GTR+Γ to have roughly equivalent cross-validation scores to those of CAT-GTR or CAT-GTR+Γ varied but was always between $f = 20$ and $f = 60$.

Empirical Mixture Models With and Without Γ

The final models displayed in figure 5 are the empirical mixture models C20-WAG, C20-Poisson, C40-WAG, C40-Poisson, C60-WAG, C60-Poisson, with and without their gamma-distributed rates variants. All of the empirical mixture models followed a similar pattern. The worst performing of any empirical mixture model set (C20, C40 and C60) was always the C20 version. The model with the next highest cross-validation score was the C40 variant in each set. Finally the C60 variants performed best of the three. This trend holds regardless of any other aspects of the model. In other words, regardless of the context, C40 outperformed C20, and C60 outperformed C40. This indicates that the richest empirical mixture model better captures across-site heterogeneity than the more compact empirical mixtures, as observed in the original work by Le et al. (2008a).

The configuration with the least complexity are the Poisson substitution process models without gamma-distributed rates-across-sites (C20-Poisson, C40-Poisson, and C60-Poisson). These versions were observed to have cross-validation scores equal to or below the GTR

model. In other words, given the choice between free exchangeability parameters or an empirical mixture model of Le et al. (2008a), the free exchangeabilities are preferred.

The next best performing set of empirical mixture models were those combined with free exchangeabilities but no Γ (C20-GTR, C40-GTR and C60-GTR). Performing even better than the set with free exchangeabilities was the set with Poisson exchangeabilities and gamma-distributed rates-across-sites (C20-Poisson+Γ, C40-Poisson+Γ, and C60-Poisson+Γ). This result is significant because it shows that addition of gamma-distributed rates-across-sites has a higher impact on how well these empirical models are able to fit the data than addition of free exchangeabilities. This contrasts with results observed under infinite mixture models (CAT and CAT-GTR), where the introduction of free exchangeabilities has a higher impact than gamma-distributed rates-across-sites. The finite empirical mixture models proposed by Le et al. (2008a) assumed even exchangeabilities. The frequency profiles that make up C20, C40 and C60 do not appear to harmonize well with variable exchangeabilities.

The top performing set of empirical mixture models was the permutation which included free exchangeabilities and Γ (C20-GTR+Γ, C40-GTR-Γ, C60-GTR+Γ). The C40-GTR+Γ and C60-GTR+Γ models performed slightly better than CAT-GTR (without Γ). In this context, all three empirical mixtures perform better than the finite mixture model variants with the same number of categories (C20-GTR+Γ was better than CAT$_f$-GTR with $f = 20$ components, C40-GTR+Γ was better than CAT$_f$-GTR with $f = 40$ components and C60-GTR+Γ was better than CAT$_f$-GTR with $f = 60$ components) although showing lower cross-validation scores than the CAT$_f$-GTR+Γ model with the same number of components. The best performing empirical mixture model set also did not have higher cross-validation

scores than CAT-GTR+Γ.

The second best set of empirical mixture models (C20-Poisson+Γ, C40-Poisson+Γ, and C60-Poisson+Γ), performed well below GTR+Γ. In other words, given the choice between free exchangeability parameters or an empirical mixture of amino acid profiles, the former is preferred. Inclusion of the C20, C40 and C60 empirical mixtures into the GTR and GTR+Γ models resulted in a improvement in model fit. C20-GTR, C40-GTR, and C60-GTR performed better than GTR and C20-GTR+Γ, C40-GTR+Γ, and C60-GTR+Γ performed better than GTR+Γ. These results are similar to those of Le et al. (2008a) who showed that empirical mixtures in a model models perform better than the homogeneous version of that model.

In all models analyzed the addition of gamma-distributed rates-across-sites saw a improvement in model fit, this is consistent with previous research on the use of gamma-distributed rates-across-sites (Yang, 1996).

Exceptions

Most of these results discussed for the Broughton-61-19997 data set, follow the same trend for the other data sets. There are, however, some exceptions.

Cross-validation analysis on the Lartillot-78-15117 data set resulted in GTR+Γ having a higher score than CAT-Poisson and CAT-Poisson+Γ. The data set Lartillot-78-15117 is one of placental mammals and highlights that the introduction of a complex mixture modelling approach, such as CAT, may not always provide the highest model fit, especially when important factors, such as amino acid exchangeabilities, are ignored. In the Lartillot-78-15117 data set, CAT-GTR+Γ still retained the highest overall cross-validation score.

In the Wainwright-188-8439 data set, the standard deviation for the cross-validation analysis of CAT-GTR+Γ was very large. It is unclear why this data set had high variability in cross-validation score, but the trend was consistent across three separate replicates. In this data sets, the cross-validation score at which CAT$_f$-GTR+Γ plateaued at was lower than the score attained by CAT-GTR+Γ. This unusual variability did not appear when analysis of these models was done without use of Γ. We suspect this data set have low evolutionary signal, but further research is required to better understand this behaviour.

Finally, when analysis of the Lartillot-78-15117 using the empirical mixture models the results were incongruent with the results in the other two data sets. The least complex empirical mixture models (C20-Poisson, C40-Poisson and C60-Poisson) performed much worse than GTR, with cross-validation scores roughly 1000 below GTR. The order of improvement for the empirical mixture models were also different for the Lartillot-78-15117 data set. Addition of gamma-distributed rates-across-sites saw less improvement in the models than did the addition of GTR exchangeabilities. The magnitude of the difference in cross-validation score was small between the C20-GTR, C40-GTR, and C60-GTR models and C20-Poisson+Γ, C40-Poisson+Γ, and C60-Poisson+Γ models (119.6, 113.2, and 73.4 respectively). For this data set it appears that use of gamma-distributed rates-across-sites is much less important than use of free exchangeabilities in phylogenetic analysis, although use of both has the highest cross-validation score regardless of the type of model used.

4. Conclusions and Future Directions

For all data sets analyzed, CAT-GTR+Γ was always the top performing model. The cross-validation score attained by CAT-GTR+Γ was often matched by CAT$_f$-GTR+Γ, with a sufficient numbers of components, typically between $f = 20$ and $f = 60$, depending on the data set.

For some data sets (Broughton-61-19997, Lartillot-78-15117), the empirical mixture model C60-GTR+Γ definitively outperform the infinite mixture model CAT+Γ and the other (Wainwright-188-8439) it performed roughly as well as CAT+Γ. It is noteworthy that C60-GTR+Γ never attained a cross-validation score comparable to CAT-GTR+Γ or CAT$_f$-GTR+Γ. Considering that the C60 frequency profile set assumed equal amino acid exchangeabilities, a new set of empirical mixture models could be created where the components of the finite mixture are estimated jointly with exchangeabilities. Our results suggest that this theoretical empirical mixture could have a high degree of model fit on amino acid data, while enjoying the computational speed of finite mixture models.

The present study does not address the potential impact of modelling across-site amino acid heterogeneity on phylogenetic inference *per se*. The current literature includes many instances in which GTR+Γ and CAT-GTR+Γ produce fundamentally different topologies (e.g., Lartillot et al., 2007; Feuda et al., 2017). Most of the differences in topologies have been attributed to the long branch attraction artifact (Felsenstein, 1978). It is also possible to modify branch lengths to induce long branch attraction artifacts in simulated data. By utilizing our systematic scan of mixture models on such data sets, both real and simulated, we should uncover the level of heterogeneity that must be recognized by a model to suppress

long branch attraction.

This analysis was done wholly in amino acid space. Repetition of this analysis, with the absence of empirical mixture models, in nucleotide space and codon state space could prove valuable. Analysis on nucleotide data would be beneficial in confirming that CAT-GTR+Γ is still the top performing model. It would also be useful to quantify how much impact gamma-distributed rates-across-sites had on model fit on different data types, and to determine how many components are required for CAT$_f$-GTR+Γ to achieve cross-validation scores equivalent to CAT-GTR+Γ. Analysis of the relative merit of empirical mixture models versus CAT-like models in codon state space would be valuable to quantify how finite and infinite mixture models perform in the data rich codon environment.

Five fold, five replicate cross-validation were used as the test settings for model comparison. Although there was no empirical justification for these particular settings, the results were significant, having adequate resolution between model cross-validation scores. Doing similar analyses as these, under different partitioning schemes would be a valuable method for determining the optimal settings required for Bayesian cross-validation analysis.

Recognizing across-site heterogeneity in amino acid profiles has also been explored within models that operate in a codon (nucleotide triplet) state space, applying the mutation-selection principle (Rodrigue et al., 2010; Rodrigue and Lartillot, 2014). These models all rely on the infinite mixture model paradigm. Significant computational improvements could be achieved by establishing a suitable finite mixture version of these models. Applying the methodology used in this study within the codon substitution framework could help establish such a finite mixture.

Altogether this study emphasized the usefulness of Bayesian cross-validation in phyloge-

netics and highlight the robustness of CAT-GTR+Γ to adapt to all data contexts.

References

Akaike, Hirotugu. 1974. A new look at the statistical model identification. *IEEE transactions on automatic control* 19:716–723.

Broughton, RE., R. Betancur-R, C. Li, G. Arratia, and G. Orti. 2013. Multi-locus phylogenetic analysis reveals the pattern and tempo of bony fish evolution. *PLOS Currents Tree of Life.* 5(1).

Echave, Julian, Stephanie J Spielman, and Claus O Wilke. 2016. Causes of evolutionary rate variation among protein sites. *Nature Reviews Genetics* 17:109.

Efron, Bradley. 2005. Bayesians, frequentists, and scientists. *Journal of the American Statistical Association* 100:1–5.

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

Felsenstein, J. 1981. Evolutionary trees from dna sequences: a maximum liklihood approach. *J. Mol. Evol.* 17:368–376.

Feuda, Roberto, Martin Dohrmann, Walker Pett, Hervé Philippe, Omar Rota-Stabelli, Nicolas Lartillot, Gert Wörheide, and Davide Pisani. 2017. Improved modeling of compositional heterogeneity supports sponges as sister to all other animals. *Current Biology* 27:3864–3870.

Guindon, S., and Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* 52:696–704.

Hastings, W Keith. 1970. Monte carlo sampling methods using markov chains and their applications .

Jukes, Thomas H, Charles R Cantor, et al. 1969. Evolution of protein molecules. *Mammalian protein metabolism* 3:132.

Kainer, D., and R. Lanfear. 2015. The effects of partitioning on phylogenetic inference. *Mol. Biol. Evol.* 32:1611–1627.

Larget, Bret, and Donald L Simon. 1999. Markov chain monte carlo algorithms for the bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.* 16:750–759.

Lartillot, N., H. Brinkmann, and H. Philippe. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a dite-heterogeneous model. *BMC Evol. Biol.* 7, Supplement 1:S4.

Lartillot, N., and F. Delsuc. 2012. Joint reconstruction of divergence times and life-history evolution in placental mammals using a phylogenetic covariance model. *Evolution.* 66-6:1773–1787.

Lartillot, N., and H. Philippe. 2004. A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* 21(6):1095–1109.

Lartillot, N., and H. Philippe. 2006. Computing bayes factors using thermodynamic integration. *Syst. BIol.* 55(2):195–207.

Lartillot, Nicolas, Thomas Lepage, and Samuel Blanquart. 2009. Phylobayes 3: a bayesian

software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* 25:2286–2288.

Le, Si Quang, and Olivier Gascuel. 2008. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* 25:1307–1320.

Le, SQ., O. Gascuel, and N. Lartillot. 2008a. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* 24(20):2317–2323.

Le, SQ., N. Lartillot, and O. Gascuel. 2008b. Phylogenetic mixture models for proteins. *Phil. Trans. R. Soc. B.* 363:3965–3976.

Li, Shuying, Dennis K Pearl, and Hani Doss. 2000. Phylogenetic tree construction using markov chain monte carlo. *Journal of the American Statistical Association* 95:493–508.

Metropolis, Nicholas, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. 1953. Equation of state calculations by fast computing machines. *The journal of chemical physics* 21:1087–1092.

Pagel, M., and A. Meade. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst. Biol.* 53(4):571–581.

Rodrigue, N., H. Philippe, and N. Lartillot. 2010. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proc. Natl. Acad. Sci. U.S.A.* 107(10):4629–4634.

Rodrigue, Nicolas, and Nicolas Lartillot. 2012. Monte carlo computational approaches in bayesian codon substitution modeling. *Codon Evolution* 45–59.

Rodrigue, Nicolas, and Nicolas Lartillot. 2014. Site-heterogeneous mutation-selection models within the phylobayes-mpi package. *Bioinformatics* 30:1020–1021.

Shapiro, Beth, Andrew Rambaut, and Alexei J Drummond. 2005. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Mol. Biol. Evol.* 23:7–9.

Stone, Mervyn. 1977. An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society. Series B (Methodological)* 44–47.

Sullivan, Jack, and Paul Joyce. 2005. Model selection in phylogenetics. *Annu. Rev. Ecol. Evol. Syst.* 36:445–466.

Susko, Edward, Léa Lincker, and Andrew J Roger. 2018. Accelerated estimation of frequency classes in site-heterogeneous profile mixture models. *Mol. Biol. Evol.* 35:1266–1283.

Tavaré, S. 1986. Some probabilistic and statistical problems in the analysis of dna sequences. *Lectures on mathematics in the life sciences* 17(2):57–86.

Wainwright, P., W. Smith, S. Price, K. Tang, J. Sparks, L. Ferry, K. Kuhn, R. Eytan, and T. Near. 2012. The evolution of pharyngognathy: A phylogenetic and functional appraisal of the pharyngeal jaw key innovation in labroid fishes and beyond. *Syst. Biol.* 61(6):1001–1027.

Wang, H., K. Li, E. Suskom, and A. 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):doi:10.1186/1471–2148–8–331.

Whelan, S., and N. Goldman. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* 18:691–699.

Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from dna sequences when substitution rates differ over sites. *Mol. Biol. Evol.* 10(6):1396–1401.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* 39(3):306–314.

Yang, Ziheng. 1996. Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology & Evolution* 11:367–372.

Yang, Ziheng, and Bruce Rannala. 1997. Bayesian phylogenetic inference using dna sequences: a markov chain monte carlo method. *Mol. Biol. Evol.* 14:717–724.