Bayesian analysis of amino acid substitution models

John P. Huelsenbeck1,*, Paul Joyce2, Clemens Lakner3 and Fredrik Ronquist4

1Department of Integrative Biology, University of California, Berkeley, 3060 VLSB #3140, Berkeley, CA 94720-3140, USA
2Departments of Mathematics and Statistics, University of Idaho, Moscow, ID 83844, USA
3School of Computational Science, DSL 150-J, Florida State University, Tallahassee, FL 32306-4120, USA
4Swedish Museum of Natural History, Box 50007, 104 05 Stockholm, Sweden

Models of amino acid substitution present challenges beyond those often faced with the analysis of DNA sequences. The alignments of amino acid sequences are often small, whereas the number of parameters to be estimated is potentially large when compared with the number of free parameters for nucleotide substitution models. Most approaches to the analysis of amino acid alignments have focused on the use of fixed amino acid models in which all of the potentially free parameters are fixed to values estimated from a large number of sequences. Often, these amino acid models are specific to a gene or taxonomic group (e.g. the Mtma model, which has parameters that are specific to mammalian mitochondrial gene sequences). Although the fixed amino acid models succeed in reducing the number of free parameters to be estimated—indeed, they reduce the number of free parameters from approximately 200 to 0—it is possible that none of the currently available fixed amino acid models is appropriate for a specific alignment. Here, we present four approaches to the analysis of amino acid sequences. First, we explore the use of a general time reversible model of amino acid substitution using a Dirichlet prior probability distribution on the 190 exchangeability parameters. Second, we then explore the behaviour of prior probability distributions that are ‘centred’ on the rates specified by the fixed amino acid model. Third, we consider a mixture of fixed amino acid models. Finally, we consider constraints on the exchangeability parameters as partitions, similar to how nucleotide substitution models are specified, and place a Dirichlet process prior model on all the possible partitioning schemes.

Keywords: phylogeny; Bayesian analysis; Markov chain Monte Carlo; Dirichlet process prior

1. INTRODUCTION

The statistical phylogenetic analysis of amino acid data presents problems that are not associated with nucleotide data. The instantaneous rate matrix is $20 \times 20$ in dimension for amino acid data. For the most general model of amino acid substitution, a model that may not be time reversible, there are a total of $20 \times 19 = 380$ parameters to be estimated. For a general time reversible (GTR) model of amino acid substitution, isomorphic to the GTR model used in phylogenetic analysis of DNA sequences (Tavaré 1986), there would be a total of $20 \times 190 = 3800$ parameters. Unlike the case with codon models in which many rates between states may be zero because the changes involve two or more substitutions in an instant of time, there are no easy ways to reduce the number of parameters to be estimated for amino acid data. The most common approach used for the phylogenetic analysis of amino acid sequence data is to use a continuous-time Markov model in which all of the rates are fixed to specific values. There are a number of such fixed rate matrices that have rates based on the analysis of inferred amino acid changes from various databases. These models include the Jones (Jones et al. 1992), Dayhoff (Dayhoff et al. 1978), Mtrev (Adachi & Hasegawa 1996), WAG (Whelan & Goldman 2001), Mtmam (Cao et al. 1998; Yang et al. 1998), Rtre (Dinhm et al. 2002), Cprev (Adachi et al. 2000), Blosum (Henikoff & Henikoff 1992), ECM (Empirical Codon Model; Kosiol et al. 2007) and Vt (Muller & Vingron 2000). The Poisson model, which is isomorphic to the Jukes & Cantor (1969) nucleotide substitution model, can also be considered a member of the family of fixed amino acid models (Bishop & Friday 1987).

The fixed amino acid models are useful not only because they reduce the number of parameters to be estimated in a phylogenetic analysis, but also because they can be applied to small alignments (datasets involving a small number of taxa and sites). It is unlikely that the rates of substitution for an amino acid model that had the rates of change free to vary could be reliably estimated for typical (small) amino acid alignments. However, the use of fixed amino acid models also complicates matters because, for any specific alignment of amino acid sequences, it is not clear which of the many potential models is the most appropriate. Often, one can make a good guess of...
which model should be used; for example, if the alignment is of plastid genes, then the Cprev model (Adachi et al. 2000) might be appropriate because its rates are based on a database of plastid genes. Similarly, the Mtmam model is probably the most appropriate for an alignment of mammalian mitochondrial genes. Yet, there is no guarantee that any specific amino acid model is the most appropriate for a particular alignment, even in cases where the amino acid model is based upon a database of genes similar to the one to be analysed. Another approach is to use the fixed model that has the maximum likelihood. This is sensible, but involves optimizing likelihoods under the current models. Also, this approach does not allow one to spread one’s bets across amino acid models if several of the models have similar likelihoods.

Even the best-fitting fixed amino acid model, however, may not be particularly suitable for the data at hand. It is interesting to note that biologists who adopt the approach of using fixed amino acid substitution models are, in a sense, adopting a Bayesian perspective, even if they do not use Bayesian methodology to estimate the parameters of the phylogenetic model. The fixed amino acid model that is assumed in the analysis can be considered a prior probability distribution on the rates of amino acid substitution for the data at hand. In fact, by using a model of amino acid substitution in which all of the rates are fixed to specific values, the biologist has adopted the strongest form of a prior that can be imagined; a fixed amino acid model places a point mass probability on the rates specified by the model, but zero probability on other rate combinations, even those rates that are slightly different from the fixed rates. An intermediate solution, in which the assumptions of the fixed amino acid model are tempered, might be more appropriate.

In this paper, several Bayesian approaches to the analysis of amino acid models are developed. We consider (i) inference of amino acid rates under a GTR model, (ii) inference of amino acid data when the rates of substitution have been ‘centred’ on a fixed amino acid model, but which still allow rates to vary, (iii) a model averaging approach, in which the results of a phylogenetic analysis are averaged over a candidate set of fixed amino acid models, and (iv) a model averaging approach in which all possible partitions of the exchangeability/rate parameters are considered.

2. MATERIAL AND METHODS

(a) Specifying a ‘centred’ prior distribution for amino acid substitution rates

We adopt a Bayesian perspective to statistical estimation, in which inferences are based on the posterior probability distribution of a parameter, which can be calculated using Bayes’ theorem as

\[ P(\text{Parameter}|\text{Data}) = \frac{P(\text{Data}|\text{Parameter})P(\text{Parameter})}{P(\text{Data})} \]

where \( P(\text{Parameter}|\text{Data}) \) is the posterior probability distribution of the parameter; \( P(\text{Data}|\text{Parameter}) \) is the likelihood; \( P(\text{Parameter}) \) is the prior probability distribution of the parameter; and \( P(\text{Data}) \) is the marginal likelihood, obtained by summing and/or integrating over all possible combinations of the model parameters. We consider the observations to be an alignment of \( s \) amino acid sequences each \( n \) in length, denoted \( X \). (We ignore the possibility that the amino acid sequences might be misaligned.) For a phylogenetic model, the parameters include a tree, with branch lengths specified in terms of expected number of substitutions per site, and a continuous-time Markov model describing how the characters, in our case amino acid sequences, change over time. We will forgo a thorough treatment of the phylogenetic model which would include a description of how the likelihood is calculated; suffice it to say that we use standard methods for calculating the likelihood (Felsenstein 1981) and use Markov chain Monte Carlo (MCMC) to numerically calculate the posterior probability distribution of the parameters (e.g. Larget & Simon 1999). Moreover, we accommodate among-site rate variation by assuming that the rate at a particular amino acid site (column in the alignment) is a random variable drawn from a mean-one gamma distribution with parameter \( \gamma \) (Yang 1993, 1994).

We assume that amino acid substitutions occur according to a continuous-time Markov model with instantaneous rates of change described by the following rate matrix

\[ Q = \{q_{ij}\} \]

which corresponds to a GTR model of amino acid substitution. (The dots represent rows and columns that are not shown because the rate matrix is too large to be printed in its entirety.) The amino acid substitution model has 20 states, corresponding to the 20 possible amino acids: \( S = \{A,R,N,D,C,Q,E,G,H,I,L,K,M,F,P,S,T,W,Y,V\} \) (which are the IUPAC codes for the amino acids; Cornish-Bowden 1985). The diagonal elements of the rate matrix are specified such that each row sums to 0. The GTR model of amino acid substitution has a total of 210 parameters. Twenty of the parameters are the stationary frequencies of the amino acids. Similar to other phylogenetic models, the stationary frequencies of the process are parameters of the model, specified as components of the rate matrix. The other 190 parameters are rate parameters, sometimes referred to as exchangeability parameters; the exchangeability parameters are the rate factors \( \theta_{AR}, \theta_{AN}, \theta_{AD}, \ldots, \theta_{VW} \).

The 20 amino acid frequencies, \( \pi = (\pi_A, \pi_R, \pi_N, \pi_D, \ldots, \pi_V) \) are, of course, constrained to sum to one. The 190 exchangeability parameters, \( \theta = (\theta_{AR}, \theta_{AN}, \theta_{AD}, \ldots, \theta_{VW}) \), on the other hand, are ideally measured in terms of the expected number of substitutions per unit time, and would not have any constraint (other than the common-sense one that the rates be positive). However, without any reference to time on a tree, the values of the rate parameters cannot be estimated; only divergence—the product of substitution rate and time—can be measured on a phylogenetic tree in the absence of a calibration. Because only divergence on a tree can be estimated, only the relative substitution rates can be estimated. For example, the relative substitution rates \( \theta_{AR} = 1, \theta_{AN} = 3, \theta_{AD} = 2, \ldots \) are equivalent to \( \theta_{AR} = 2, \theta_{AN} = 6, \theta_{AD} = 4, \ldots \). In this study, we impose the constraint that the 190 rate parameters sum to one. The rate matrix is scaled by a factor
\[ f(x) = \frac{1}{b(a)} \prod_{i=1}^{K} x_i^{a_i - 1}, \]

where \( a = (a_1, a_2, \ldots, a_K) \) are the parameters of the distribution and the constant of integration is a well-known ratio of gamma functions given by

\[ b(a) = \frac{\prod_{i=1}^{K} \Gamma(a_i)}{\Gamma(\sum_{i=1}^{K} a_i)}. \]

The marginal probability distribution of the Dirichlet random variable is a beta distribution with parameters \( a_i \) and \( a_0 - a_i \). The expected value for the \( i \)-th Dirichlet random variable is \( E(X_i) = a_i/a_0 \) and the variance is \( \text{var}(X_i) = (a_i(a_0 - a_i))/a_0^2(a_0 + 1) \).

We assume that the amino acid frequencies follow a flat Dirichlet prior distribution in which \( a_1 = a_2 = \ldots = a_20 = 1 \) (\( K = 20 \)). We consider several different prior distributions for the exchangeability parameters, \( \theta \). In general, the exchangeability parameters have the following Dirichlet probability distribution:

\[ f(\theta | \phi) = \frac{1}{b(\phi)} \prod_{i < j} \theta_{ij}^{\phi_{ij} - 1}, \]

where \( S \) is the set of ordered amino acid states. The constant of integration is again, a ratio of gamma functions. The parameters \( \phi_{ij} \) are constrained to sum to one. We interpret the parameter \( \phi \) as a concentration parameter that controls the variance of the Dirichlet prior distribution. The expectation and variance of the exchangeability parameters are

\[ E(\theta_{ij}) = \phi_{ij}/\phi \quad \text{and} \quad \text{var}(\theta_{ij}) = (\phi_{ij}(1 - \phi_{ij})/\phi + 1), \]

Note that the maximum variance allowed is \( \phi_{ij} = 1 \) that occurs at the improper prior \( \phi = 0 \) and is the variance associated with a single draw from a binomial with proportion \( \phi_{ij} \). One also show that cov\( (\theta_{ij}, \theta_{km}) = -\phi_{ij}(1 - \phi_{ij})/\phi + 1 \) \( \theta_{ij} \) and \( \theta_{km} \) will be most correlated when \( \phi = 0 \) which corresponds to the covariance associated with a single draw from a multinomial \( (\phi_{ij}, \phi_{km}, 1 - \phi_{ij} - \phi_{km}) \).

We consider Dirichlet prior distributions for the exchangeability parameters that are centred on different values. To centre the prior distribution on a particular fixed amino acid rate matrix, we set the \( \phi_{ij} \) such that they are equal to the scaled entries of the fixed rate matrix. (A similar approach of using an informative Dirichlet prior was described by Zwickl & Holder 2004.) Consider, for example, the following fixed rate matrix with only three states

\[ Q = \begin{pmatrix}
- & 1.0 \pi_1 & 1.2 \pi_2 \\
1.0 \pi_1 & - & 1.4 \pi_2 \\
1.2 \pi_1 & 1.4 \pi_2 & -
\end{pmatrix} \mu.
\]

A Dirichlet prior distribution centred on this rate matrix would have \( \phi_{01} = 1.0(1.0 + 1.2 + 1.4) = 0.278 \), \( \phi_{02} = 1.2/(1.0 + 1.2 + 1.4) = 0.333 \) and \( \phi_{12} = 1.4(1.0 + 1.2 + 1.4) = 0.389 \). The concentration parameter, \( \chi_i \), specifies how the probability density is spread around the centred rate value.

When \( \chi \) is small, the prior probability, while still centred on the values of the fixed rate matrix, has more prior density on values that are quite different from those specified by the rate matrix. On the other hand, large values of \( \chi \) put very little prior probability density on rates that are different from those specified by the fixed model. Note, in fact, that as \( \chi \rightarrow \infty \), the variance decreases to zero; the centred model becomes equivalent to the fixed model, with no probability density on rate values even slightly different from the fixed rates. The small example given here involves only three rates. However, the reader should see that the 3-state model is isomorphic with a 20-state amino acid model, but with

\[ \begin{pmatrix}
20 \\
2
\end{pmatrix} = 190, \]

centred rates instead of

\[ \begin{pmatrix}
3 \\
2
\end{pmatrix} = 3, \]

centred rates.

Our parametrization of the Dirichlet prior for the exchangeability parameters allows us to explore a large number of prior models on the rates. For example, we can specify the GTR model, with a flat prior, by setting all \( \phi_{ij} = 1/190 \) and \( \chi = 190 \). Similarly, we can explore the sensitivity of rate parameter estimates centred on fixed amino acid models by varying the concentration parameter \( \chi \).

(b) Model choice and model averaging

In this paper, we consider 10 amino acid models, denoted \( M_1, M_2, \ldots, M_{10} \). The 10 amino acid models correspond to the Poisson (\( M_1 \); Bishop & Friday 1987), Jones (\( M_2 \); Jones et al. 1992), Dayhoff (\( M_3 \); Dayhoff et al. 1978), Mtrev (\( M_4 \); Adachi & Hasegawa 1996), Mtmtam (\( M_5 \); Cao et al. 1998; Yang et al. 1998), WAG (\( M_6 \); Whelan & Goldman 2001), Rtrev (\( M_7 \); Doolittle et al. 2002), Cpreev (\( M_8 \); Adachi et al. 2000), Vt (\( M_9 \); Muller & Vingron 2000) and Blossom (\( M_{10} \); Henikoff & Henikoff 1992) models.

In a Bayesian analysis, inferences are based upon the joint posterior probability distribution of the model parameters. For the phylogeny problem, the joint posterior probability is calculated using Bayes’ theorem as

\[ f(\tau, \nu, M | X) = \frac{f(X | \tau, \nu, M) f(\nu | M) f(\tau | X, \nu, M)}{\int f(X | \tau, \nu, M) f(\nu | M) f(\tau | X, \nu, M) d\nu d\tau}. \]

where \( f(X | \tau, \nu, M) \) is the likelihood for the \( i \)-th tree and \( i \)-th amino acid model; \( f(\nu) \) is the prior probability density distribution for branch lengths; and \( T(s) \) is the number of unrooted trees possible for \( s \) species \( \{T(s) = (2s - 5)!!\} \). In this study, all trees and all amino acid models are considered to be equally probable, \textit{a priori}. If one is interested only in the phylogeny of the species, then inferences are based upon the marginal posterior probability distribution of phylogenetic trees. The marginal posterior probability of the \( i \)-th tree is obtained by integrating over the other model parameters:

\[ f(\tau | X) = \sum_{j=1}^{10} \left( f(\tau, \nu, M_j | X) d\nu \right). \]

Note that the estimate of phylogeny does not depend upon any particular fixed model of amino acid substitution; the posterior probability of a tree is averaged over all 10 amino acid models. This formulation of the problem does not force the biologist to choose a particular amino acid model.

Phil. Trans. R. Soc. B (2008)
The posterior probability distribution of phylogenetic trees cannot be calculated analytically. MCMC (Metropolis et al. 1953; Hastings 1970), however, can be used to approximate the posterior probability of a tree. The details of MCMC as applied to the phylogeny problem have been described elsewhere (see Larget & Simon 1999; Huelsenbeck et al. 2001). In short, the posterior probability distribution of parameters is approximated using a Markov chain which has a stationary distribution that is the posterior probability distribution of the parameters. New states for the chain (trees and branch lengths) are proposed using a stochastic mechanism and accepted or rejected according to the formula of Metropolis et al. (1953) and Hastings (1970). States are sampled from the chain when at stationarity. The fraction of the time the chain dwells on any particular tree is a valid approximation of the posterior probability of that tree. We note one important change in implementing the mixed model approximation of the posterior probability of that tree. We label possible models by introducing a notation that allows all of the possible models to be labelled. Here, there are three exchangeability parameters, and the possible labels for the models are
\[\begin{array}{ccc}
01 & 02 & 12 \\
1 & 1 & 1 \\
1 & 1 & 2 \\
1 & 2 & 1 \\
1 & 2 & 2 \\
1 & 2 & 3 \\
\end{array}\]

The first model, with label 1, 1, 1, constrains the three exchangeability parameters to be equal to one another. Similarly, the second model listed, 1, 1, 2, constrains \(\theta_{01} = \theta_{02}\), but allows \(\theta_{12}\) to vary freely (subject to the constraint that the three exchangeability parameters sum to one). Huelsenbeck et al. (2004) introduced this notation to describe all of the submodels of the GTR model of nucleotide substitution. Here, the set of exchangeability parameters is considered a partition, and the possible partitions are labelled according to the restricted growth function notation of Stanton & White (1986). The same scheme for labelling nucleotide models is implemented in the program PAUP* (Swofford 1998), but with the indices being letters instead of numbers.

The number of possible partitions of \(n\) elements is described by the Bell numbers (Bell 1934). The Bell numbers are defined as
\[B(n) = \sum_{k=0}^{n} S_2(n, k),\]
where \(S_2(n, k)\) is the Stirling number of the second kind
\[S_2(n, k) = \frac{1}{k!} \sum_{i=0}^{k-1} (-1)^i \binom{k}{i} (k-i)^n\]
and is the number of ways to partition \(n\) elements into \(k\) subsets. For the small example given above consisting of three states, the total number of ways to partition the rates is \(B(3) = 5\). Similarly, when only the exchangeability parameters are considered, there are a total of \(B(6) = 203\) possible time-reversible four-state nucleotide models (Huelsenbeck et al. 2004). (Note that there are many more possible models when stationary frequencies are considered to be fixed or estimated parameters of the model.) The number of possible time-reversible amino acid models is vast. A general time-reversible model of amino acid substitution has 190 exchangeability parameters, meaning that there are a total of \(B(190) = 6.59 \times 10^{258}\) possible models, the most parameter rich of which has 190 independently estimated rates and the simplest of which has a common rate for all 190 exchangeability parameters.

We consider phylogenetic analysis of amino acid data when the candidate pool of substitution models consists of all \(6.59 \times 10^{258}\) time-reversible substitution models. Huelsenbeck et al. (2004) performed analysis of nucleotide data on the set of 203 possible time-reversible nucleotide models, placing a uniform prior probability distribution on all possible models. However, directly extending the approach of Huelsenbeck et al. (2004) to the amino acid data is not feasible for several reasons. First, if a uniform prior probability distribution is placed on all possible amino acid models, the result is a prodigious amount of probability on models with an intermediate number of substitution rates. For example, models

\[\begin{array}{ccc}
01 & 02 & 12 \\
1 & 1 & 1 \\
1 & 1 & 2 \\
1 & 2 & 1 \\
1 & 2 & 2 \\
1 & 2 & 3 \\
\end{array}\]

\(\theta_{01} = \theta_{12}\) and \(\theta_{02} = \theta_{12}\). The last possible model does not constrain equalities among the exchangeability parameters. We label possible models by introducing a notation that allows all of the possible models to be labelled. Here, there are three exchangeability parameters, and the possible labels for the models are

\[\begin{array}{ccc}
01 & 02 & 12 \\
1 & 1 & 1 \\
1 & 1 & 2 \\
1 & 2 & 1 \\
1 & 2 & 2 \\
1 & 2 & 3 \\
\end{array}\]
with \( k = 48 \) substitution types have a probability of \( S_2(190, 48)/B(190) = 0.133 \), whereas models with only \( k = 3 \) rate classes have a probability of \( S_2(190, 3)/B(190) = 1.14 \times 10^{-169} \). The posterior probability of the number of substitution types is strongly affected by the prior, and is drawn to models with approximately 50 rate classes simply because there are so many more models with an intermediate number of rate classes. Second, the MCMC used in the Huelsenbeck et al. (2004) paper does not work efficiently for a problem with 190 substitution types.

We take a different approach in this paper. We use a Dirichlet process prior model (Ferguson 1973; Antoniak 1974) as a prior probability on substitution models. The Dirichlet process prior is a prior model on partitions and has been usefully applied in several contexts in phylogenetics and evolutionary biology (Lartillot & Philippe 2004; Huelsenbeck et al. 2006). Moreover, robust MCMC methods have been developed for this model, which allow effective exploration of the space of partitions (Neal 2000). The Dirichlet process prior has a single parameter, here denoted \( \beta > 0 \), that controls the ‘clumpiness’ of the process. A simple description of the Dirichlet process prior is as follows. Consider adding rate classes sequentially. Let the total number of rate classes be \( K = \sum_{i=1}^{n} Y_i \). Here \( n \) is the total number of possible rates \( (n = 190) \). Therefore, \( E(K) = \beta \sum_{i=1}^{n} \frac{1}{1+i} = \beta \ln(n) + \beta \). Small values for \( \beta \) result in only a few groups of substitution types \( (k) \) whereas large values of \( \beta \) place more probability on large values of \( k \). In fact, the prior probability of two rate parameters being placed in the same group is \( 1/1+\beta \). Since the number of substitution types is on the order of the natural logarithm of the total number of substitutions, the prior gives more weight to fewer classes than the uniform prior on partitions.

(d) Data
We analysed eight alignments of amino acid sequences: (i) \( adh \) sequences from \( s = 23 \) Drosophila species (Cao et al. 1998; Yang et al. 2000), (ii) \( \beta \)-globin sequences from \( s = 17 \) vertebrates (Yang et al. 2000), (iii) coat protein sequences from \( s = 9 \) viruses from the family Leviriviridae (Bollback & Huelsenbeck 2001), (iv) \( env \) sequences from \( s = 23 \) Japanese encephalitis viral samples (Yang et al. 2000), (v) \( pol \) sequences from \( s = 23 \) HIV samples (Yang et al. 2000), (vi) \( i = 28 \) hemagglutinin sequences from influenza virus (type A; Fitch et al. 1997; Yang et al. 2000), (vii) replicase sequences from \( s = 9 \) viruses from the family Lepiriviridae (Bollback & Huelsenbeck 2001) and (viii) E-glycoprotein sequences sampled from \( s = 18 \) flavirvirus (Zanotto et al. 1996; Yang et al. 2000).

3. RESULTS
For each analysis of this paper, two MCMC analyses, each of four million cycles, were performed. Paired chains were checked for convergence to the same marginal probability distribution for the 190 exchangeability, 20 state frequency and gamma-shape parameters using the program Tracer (Rambaut & Drummond 2007) and boa (Smith 2007). Inferences were based on samples taken after the one millionth MCMC cycle. Analysis of the post-burn-in samples taken by each pair of chains shows that all analyses had estimated sample sizes greater than 100 (for the combined analyses) and resulted in very similar (by eye) marginal posterior probability distributions for all of the exchangeability parameters. We also calculated the ‘potential scale reduction factor’ statistic \( \sqrt{\hat{R}} \) of Gelman & Rubin (1992) using the program boa (Smith 2007). The potential scale reduction compares the between-chain variance to the within-chain variance and approaches one from above. Values of \( \sqrt{\hat{R}} \) less than 1.2–1.3 are taken as evidence that the chain has converged for that parameter (Gelman 1996). Figure 1 shows the values of \( \sqrt{\hat{R}} \) for all of the parameters shown in this paper. The vast majority of these parameters have \( \sqrt{\hat{R}} \) less than 1.05, with only a few having values of \( \sqrt{\hat{R}} \) between 1.2 and 1.3.

(a) Analysis under the GTR model of amino acid substitution
We estimated the exchangeability parameters of the GTR model of amino acid substitution under a flat Dirichlet prior probability distribution. In our formulation, the flat Dirichlet distribution has all \( v_{ij} = 1/190 \) and \( \chi = 190 \). We used MCMC to approximate the joint posterior probability distribution of the exchangeability parameters \( \theta \), amino acid frequency parameters \( \pi \), gamma-shape parameter \( \gamma \), phylogenetic tree \( \tau \) and branch-length parameters \( \nu \).

Figure 2 shows the marginal posterior and prior probability distributions of four of the exchangeability parameters for the HIV alignment. The marginal posterior probability distribution for a parameter integrates over uncertainty in all of the other model parameters. Hence, the marginal distributions shown in figure 2 are not conditioned on any particular phylogenetic tree, set of branch lengths, etc., but rather account for uncertainty in these nuisance parameters. The prior probability distribution of a single exchangeability parameter follows a beta distribution with parameters 1 and 189. The posterior distribution differs for each of the parameters. The posterior probability distribution for the rate of substitution between amino acids \( M \) and \( Y \) closely resembles the
Figure 2. The marginal prior and posterior probability distribution for four of the exchangeability parameters for the HIV pol alignment. The Kullback–Leibler divergence, \( I \), is shown for each parameter and measures the dissimilarity between the prior and posterior probability distributions. Note that the Kullback–Leibler divergence is small for many of the exchangeability parameters. Note that the Kullback–Leibler divergence can be interpreted as a measure of the informativeness of the data about the value of the exchangeability parameter.

The amino acid sequence data are informative for some, but not all, of the exchangeability parameters. Figure 3 shows a summary for each of the eight alignments examined in this study. For each alignment, the mean of the posterior distribution and the Kullback–Leibler divergence is shown for the 190 exchangeability parameters. Note that the Kullback–Leibler divergence is small for many of the exchangeability parameters. In these cases, the data are not particularly informative about the value of the parameter. However, in some cases, the Kullback–Leibler divergence is large. The data are informative about the realized value of these exchangeability parameters.

(b) Analysis under ‘centred’ fixed amino acid prior models

We analysed one of the eight alignments (the HIV alignment) under four ‘centred’ prior models. Specifically, we analysed the data under a prior centred on the Poisson (Bishop & Friday 1987), Blosom (Henikoff & Henikoff 1992), Jones (Jones et al. 1992) and WAG (Whelan & Goldman 2001) models. These analyses were designed to address three questions: (i) How are the substitution rates affected by the prior model? (ii) How informative is the data about the substitution rates? and (iii) What is the effect of the concentration parameter, \( \chi \), on the exchangeability parameters?

Figure 4 shows the marginal prior and posterior probability densities for a few of the substitution rates. Specifically, figure 4 shows the marginal prior and posterior probability densities of the \( I \rightarrow V \) (\( \theta_{RV} \)) and \( M \rightarrow F \) (\( \theta_{MF} \)) parameters for the HIV alignment. In each case, the exchangeability parameters were chosen to display an instance in which the data are informative (\( I \rightarrow V \)) and rather uninformative (\( M \rightarrow F \)) about the parameter’s value. Note that the posterior distribution is shifted away from the prior distribution for the case in which the data are informative. This shift towards larger values of the exchangeability parameter results in a larger Kullback–Leibler divergence when compared with the case in which the data are uninformative. The prior and posterior probability densities are almost indistinguishable for the uninformative rate, resulting in a small Kullback–Leibler divergence. Figure 5 shows estimated from the MCMC output. The Kullback–Leibler divergence between two beta distributions, one with parameters \( a_1 \) and \( b_1 \) and the other with parameters \( a_2 \) and \( b_2 \) is

\[
I = \ln \left( \frac{b(a_2, b_2)}{b(a_1, b_1)} \right) - (a_2 - a_1) \psi(a_1) - (b_2 - b_1) \psi(b_1)
\]

\[
+ (a_2 - a_1 + b_2 - b_1) \psi(a_1 + b_1),
\]

where, as earlier, the beta function \( b(x, y) = I(x) I(y)/I(x + y) \) is the ratio of gamma functions and \( \psi(x) = I'(x)/I(x) \) is the digamma function. Figure 2 shows the Kullback–Leibler divergence between the prior and posterior distributions for the exchangeability parameters, \( R \rightarrow K \), \( I \rightarrow V \), \( N \rightarrow D \) and \( M \rightarrow Y \). Note that the Kullback–Leibler divergence is large when the prior and posterior distributions are dissimilar. The Kullback–Leibler divergence can be closely approximated with parameters that are
the Kullback–Leibler divergences for all 190 exchangeability parameters for the HIV alignment. In general, the posterior probability distribution of an exchangeability parameter becomes more similar to the prior distribution as the concentration parameter, $c$, increases.

The concentration parameter, $c$, strongly influences the posterior probability distributions of the exchangeability parameters. Figure 6 shows the marginal posterior probability distribution for one of the exchangeability parameters ($I \rightarrow V$) for the HIV alignment when the prior is centred on the Poisson model. The marginal posterior distributions are most similar when $c$ is small (e.g. when $c = 0.1$ and $c = 1$), but quite different for larger values of $c$. A large value for the concentration parameter states that many prior observations of exchanges (substitutions) between amino acids were observed. Why do we say this? For most statistical problems, the Dirichlet probability distribution is used as a prior for multinomially distributed data. The Dirichlet distribution is conjugated with the multinomial distribution, which means that when a Dirichlet prior distribution is combined with a multinomial-likelihood function, the posterior distribution is also a Dirichlet distribution (but with different parameter values than the prior). The parameter values of the Dirichlet prior are interpreted as the prior number of observations. Hence, a Dirichlet with parameters $a_1 = 1, a_2 = 1, a_3 = 1$ and $a_4 = 1$ essentially assumes four prior observations (one for each of the four categories). In our use of the Dirichlet prior distribution, the $c$ parameter can be interpreted as the prior number of observations. Hence, a value of $c = 100$ might be interpreted as there being 100 prior observations. For many datasets, even a value of $c = 100$ might be considered large, and potentially swamp the information in the alignment about the values of the exchangeability parameters. Small values of $c$, however, do not appear to strongly affect the posterior distribution of substitution rate.

(c) Averaging over fixed amino acid models
We performed phylogenetic analyses under a mixture of the fixed amino acid models. The mixture model was not applied on a per site basis (i.e. the likelihood for each site is calculated as an average over the 10
fixed amino acid models), but rather each amino acid model was applied to the entire alignment (i.e. only a single amino acid model is used to calculate the likelihood for the alignment, and MCMC is used to perform the model averaging). We summarize the results of these analyses using the posterior probabilities and Bayes factors for the 10 fixed amino acid models (tables 1 and 2). In general, the $M_2$ (Jones et al. 1992), $M_3$ (Dayhoff et al. 1978) and $M_6$ (Whelan & Goldman 2001) models performed the best for the eight alignments examined in this study.

(d) Considering the substitution model as a partition of substitution rates

We performed analyses in which the substitution models were considered partitions of the 190 substitution rates. As described above, we assumed a Dirichlet process prior probability model that places some probability on all $6.59 \times 10^{236}$ possible time-reversible amino acid substitution models and explored the consequence of using different values for the concentration parameter of the Dirichlet process prior ($\beta$). Specifically, we fixed $\beta$ such that the prior mean for the number of rate categories ($K$) was 2, 5 and 10 (which are achieved by using $\beta = 0.18$, 0.81 and 2.09, respectively). Figure 7 summarizes one important aspect of the MCMC analyses: the number of rate categories explored by the Markov chain. Note that in none of the analyses was there much posterior probability on $K=1$, even when a lot of prior probability was placed on the simplest possible amino acid model, as was the case when $E(K)=2$. The posterior probability for $K=1$ was close to zero for all eight analyses and for each prior on $K$ that we explored. Although the data were informative about small values of $K$, they were less so for large values of $K$. The prior and posterior probability distributions on the number of substitution rate categories were similar when the prior mean for the number of categories was $E(K)=10$.

Even though the posterior probability distribution on the number of substitution types varied from one analysis to another, mostly depending on the prior placed on $K$, summaries of the substitution models visited were remarkably consistent. Figure 8 shows the 'mean partition' for each of the three analyses we

\[ \chi^2 = 9.2 \]
\[ \chi^2 = 4.8 \]
\[ \chi^2 = 10.4 \]
\[ \chi^2 = 82.7 \]
\[ \chi^2 = 59.9 \]
\[ \chi^2 = 5.1 \]
\[ \chi^2 = 1.9 \]
\[ \chi^2 = 3.4 \]
\[ \chi^2 = 0.171 \]
\[ \chi^2 = 0.006 \]
\[ \chi^2 = 0.114 \]
\[ \chi^2 = 0.047 \]
\[ \chi^2 = 0.019 \]
\[ \chi^2 = 0.009 \]
\[ \chi^2 = 0.001 \]
\[ \chi^2 = 0.014 \]
conducted for each alignment. The mean partition is the partition that minimizes the squared distance to all of the sampled partitions (Huelsenbeck & Andolfatto 2007). We use a distance on partitions, described by Gusfield (2002). The distance between two partitions is the minimum number of elements that must be moved between subsets to make one of the partitions identical to the other. (Or, equivalently, it is the minimum number of elements that must be deleted to make the induced partitions the same.) The mean partitions are similar under the three different choices of $\beta$ we examined for each alignment, with the same exchangeability parameters usually being grouped together.

4. DISCUSSION
Amino acid substitution models are quite complex and parameter rich with regard to both the number of substitution rates to be estimated and the number of ways to partition rates. No single dataset is expected to contain enough information to resolve the phylogeny while simultaneously providing accurate estimates of all parameters. A synthesis of various data sources is necessary, and the Bayesian approach described above provides a statistically rigorous framework that performs this synthesis. We explored several different approaches to the analysis of amino acid data, three of which are illustrated in Figure 5 and 6.

Figure 5. The Kullback–Leibler divergences ($y$-axes) of the 190 exchangeability parameters (arranged along the $x$-axes) under four different centred prior distributions for the HIV alignment.

Figure 6. The marginal posterior probability distribution of the $I\leftrightarrow V$ exchangeability parameter for the HIV alignment when $\chi$ varies. The $y$-axis is the marginal posterior probability density of the rate. Red line, 0.1; blue line, 1; light green line, 100; orange line, 190; light blue line, 500; green line, 1000; yellow line, 10 000.
Table 1. The posterior probabilities of the 10 fixed amino acid models for each of the eight alignments. (A, *Drosophila adh*; B, vertebrate β-globin; C, Leviriviridae coat; D, Japanese encephalitis *env*; E, flavivirus; F, influenza; G, HIV *pol*; H, Leviriviridae replicase.)

<table>
<thead>
<tr>
<th>model</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>M</em>1</td>
<td>$6.20 \times 10^{-101}$</td>
<td>$3.69 \times 10^{-119}$</td>
<td>$1.33 \times 10^{-66}$</td>
<td>$1.52 \times 10^{-72}$</td>
<td>$1.56 \times 10^{-151}$</td>
<td>$9.01 \times 10^{-55}$</td>
<td>$5.35 \times 10^{-294}$</td>
<td>$2.54 \times 10^{-95}$</td>
</tr>
<tr>
<td><em>M</em>2</td>
<td>$6.20 \times 10^{-3}$</td>
<td>$1.04 \times 10^{-10}$</td>
<td>$1.41 \times 10^{-8}$</td>
<td>$9.99 \times 10^{-1}$</td>
<td>$9.99 \times 10^{-1}$</td>
<td>$9.99 \times 10^{-1}$</td>
<td>$9.99 \times 10^{-1}$</td>
<td>$9.99 \times 10^{-1}$</td>
</tr>
<tr>
<td><em>M</em>3</td>
<td>$2.81 \times 10^{-11}$</td>
<td>$9.99 \times 10^{-1}$</td>
<td>$1.39 \times 10^{-8}$</td>
<td>$2.11 \times 10^{-15}$</td>
<td>$1.71 \times 10^{-18}$</td>
<td>$1.27 \times 10^{-14}$</td>
<td>$4.12 \times 10^{-45}$</td>
<td>$1.09 \times 10^{-26}$</td>
</tr>
<tr>
<td><em>M</em>4</td>
<td>$5.55 \times 10^{-34}$</td>
<td>$2.59 \times 10^{-53}$</td>
<td>$1.52 \times 10^{-60}$</td>
<td>$3.40 \times 10^{-33}$</td>
<td>$2.47 \times 10^{-87}$</td>
<td>$3.16 \times 10^{-33}$</td>
<td>$3.04 \times 10^{-131}$</td>
<td>$3.95 \times 10^{-91}$</td>
</tr>
<tr>
<td><em>M</em>5</td>
<td>$5.46 \times 10^{-45}$</td>
<td>$8.65 \times 10^{-78}$</td>
<td>$7.53 \times 10^{-89}$</td>
<td>$3.72 \times 10^{-32}$</td>
<td>$1.92 \times 10^{-113}$</td>
<td>$5.07 \times 10^{-37}$</td>
<td>$2.56 \times 10^{-143}$</td>
<td>$6.65 \times 10^{-121}$</td>
</tr>
<tr>
<td><em>M</em>6</td>
<td>$9.88 \times 10^{-1}$</td>
<td>$5.04 \times 10^{-6}$</td>
<td>$9.99 \times 10^{-1}$</td>
<td>$4.10 \times 10^{-8}$</td>
<td>$7.81 \times 10^{-6}$</td>
<td>$1.54 \times 10^{-8}$</td>
<td>$7.22 \times 10^{-31}$</td>
<td>$2.27 \times 10^{-2}$</td>
</tr>
<tr>
<td><em>M</em>7</td>
<td>$4.53 \times 10^{-3}$</td>
<td>$1.04 \times 10^{-13}$</td>
<td>$4.14 \times 10^{-9}$</td>
<td>$2.68 \times 10^{-32}$</td>
<td>$3.21 \times 10^{-20}$</td>
<td>$4.11 \times 10^{-24}$</td>
<td>$1.99 \times 10^{-12}$</td>
<td>$6.91 \times 10^{-10}$</td>
</tr>
<tr>
<td><em>M</em>8</td>
<td>$6.67 \times 10^{-3}$</td>
<td>$5.65 \times 10^{-19}$</td>
<td>$1.97 \times 10^{-9}$</td>
<td>$2.99 \times 10^{-10}$</td>
<td>$5.26 \times 10^{-7}$</td>
<td>$1.20 \times 10^{-8}$</td>
<td>$6.49 \times 10^{-17}$</td>
<td>$5.68 \times 10^{-10}$</td>
</tr>
<tr>
<td><em>M</em>9</td>
<td>$2.24 \times 10^{-7}$</td>
<td>$1.77 \times 10^{-27}$</td>
<td>$5.14 \times 10^{-5}$</td>
<td>$1.34 \times 10^{-15}$</td>
<td>$9.83 \times 10^{-17}$</td>
<td>$2.96 \times 10^{-17}$</td>
<td>$1.31 \times 10^{-44}$</td>
<td>$4.28 \times 10^{-18}$</td>
</tr>
<tr>
<td><em>M</em>10</td>
<td>$1.51 \times 10^{-10}$</td>
<td>$8.13 \times 10^{-24}$</td>
<td>$4.43 \times 10^{-3}$</td>
<td>$6.40 \times 10^{-23}$</td>
<td>$2.21 \times 10^{-20}$</td>
<td>$4.00 \times 10^{-22}$</td>
<td>$3.12 \times 10^{-67}$</td>
<td>$9.77 \times 10^{-1}$</td>
</tr>
</tbody>
</table>

Table 2. The Bayes factors of the 10 fixed amino acid models for each of the eight alignments. (A, *Drosophila adh*; B, vertebrate β-globin; C, Leviriviridae coat; D, Japanese encephalitis *env*; E, flavivirus; F, influenza; G, HIV *pol*; H, Leviriviridae replicase.)

<table>
<thead>
<tr>
<th>model</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>M</em>1</td>
<td>$5.58 \times 10^{-100}$</td>
<td>$3.32 \times 10^{-118}$</td>
<td>$1.19 \times 10^{-65}$</td>
<td>$1.41 \times 10^{-71}$</td>
<td>$1.41 \times 10^{-150}$</td>
<td>$8.11 \times 10^{-54}$</td>
<td>$4.81 \times 10^{-253}$</td>
<td>$2.29 \times 10^{-94}$</td>
</tr>
<tr>
<td><em>M</em>2</td>
<td>$5.61 \times 10^{-2}$</td>
<td>$9.41 \times 10^{-10}$</td>
<td>$1.26 \times 10^{-7}$</td>
<td>$2.17 \times 10^{-8}$</td>
<td>$1.08 \times 10^{-6}$</td>
<td>$3.26 \times 10^{-8}$</td>
<td>$4.51 \times 10^{-12}$</td>
<td>$1.52 \times 10^{-9}$</td>
</tr>
<tr>
<td><em>M</em>3</td>
<td>$2.53 \times 10^{-10}$</td>
<td>$1.78 \times 10^{9}$</td>
<td>$1.25 \times 10^{-7}$</td>
<td>$1.89 \times 10^{-14}$</td>
<td>$1.54 \times 10^{-17}$</td>
<td>$1.14 \times 10^{-13}$</td>
<td>$3.71 \times 10^{-44}$</td>
<td>$9.81 \times 10^{-26}$</td>
</tr>
<tr>
<td><em>M</em>4</td>
<td>$5.00 \times 10^{-33}$</td>
<td>$2.33 \times 10^{-52}$</td>
<td>$1.37 \times 10^{-59}$</td>
<td>$3.06 \times 10^{-32}$</td>
<td>$2.22 \times 10^{-86}$</td>
<td>$2.84 \times 10^{-32}$</td>
<td>$2.73 \times 10^{-130}$</td>
<td>$3.56 \times 10^{-90}$</td>
</tr>
<tr>
<td><em>M</em>5</td>
<td>$4.91 \times 10^{-43}$</td>
<td>$7.78 \times 10^{-77}$</td>
<td>$6.78 \times 10^{-88}$</td>
<td>$3.34 \times 10^{-31}$</td>
<td>$1.73 \times 10^{-112}$</td>
<td>$4.56 \times 10^{-36}$</td>
<td>$2.31 \times 10^{-142}$</td>
<td>$5.99 \times 10^{-120}$</td>
</tr>
<tr>
<td><em>M</em>6</td>
<td>$7.80 \times 10^{2}$</td>
<td>$4.53 \times 10^{-5}$</td>
<td>$9.39 \times 10^{4}$</td>
<td>$3.69 \times 10^{-7}$</td>
<td>$7.03 \times 10^{-5}$</td>
<td>$1.39 \times 10^{-7}$</td>
<td>$6.49 \times 10^{-30}$</td>
<td>$2.09 \times 10^{1}$</td>
</tr>
<tr>
<td><em>M</em>7</td>
<td>$4.09 \times 10^{-2}$</td>
<td>$9.44 \times 10^{-13}$</td>
<td>$3.73 \times 10^{-8}$</td>
<td>$2.41 \times 10^{-31}$</td>
<td>$2.89 \times 10^{-19}$</td>
<td>$3.70 \times 10^{-23}$</td>
<td>$1.79 \times 10^{-11}$</td>
<td>$6.22 \times 10^{-9}$</td>
</tr>
<tr>
<td><em>M</em>8</td>
<td>$6.01 \times 10^{-3}$</td>
<td>$5.09 \times 10^{-18}$</td>
<td>$1.77 \times 10^{-8}$</td>
<td>$2.69 \times 10^{-39}$</td>
<td>$4.74 \times 10^{-6}$</td>
<td>$1.08 \times 10^{-7}$</td>
<td>$5.84 \times 10^{-16}$</td>
<td>$5.12 \times 10^{-9}$</td>
</tr>
<tr>
<td><em>M</em>9</td>
<td>$2.02 \times 10^{-6}$</td>
<td>$1.60 \times 10^{-26}$</td>
<td>$4.62 \times 10^{-4}$</td>
<td>$1.21 \times 10^{-14}$</td>
<td>$8.85 \times 10^{-16}$</td>
<td>$2.66 \times 10^{-16}$</td>
<td>$1.17 \times 10^{-43}$</td>
<td>$3.85 \times 10^{-17}$</td>
</tr>
<tr>
<td><em>M</em>10</td>
<td>$1.36 \times 10^{-9}$</td>
<td>$7.32 \times 10^{-23}$</td>
<td>$3.98 \times 10^{-4}$</td>
<td>$5.76 \times 10^{-22}$</td>
<td>$1.98 \times 10^{-19}$</td>
<td>$3.60 \times 10^{-21}$</td>
<td>$2.81 \times 10^{-66}$</td>
<td>$3.86 \times 10^{2}$</td>
</tr>
</tbody>
</table>
which extend previous work on the development of fixed amino acid models, and one of which extends work on model averaging of DNA substitution models (Huelsenbeck et al. 2004).

One possible approach, explored here and implemented in MrBayes (Ronquist & Huelsenbeck 2003), is to simply average inferences over a set of fixed amino acid models. This approach has the advantage that it automates the choice of fixed amino acid models, selecting the model or models that are most appropriate for the data in hand. In practice, however, we found that little averaging occurred, because virtually all of the posterior probability is placed on a single amino acid model. Moreover, this approach assumes that one of the fixed amino acid models is appropriate for the data in hand, though in reality none

Figure 7. The prior and posterior probability distributions (upside down and right side up, respectively) for the number of exchangeability parameter groups for the analyses where the concentration parameter is fixed such that the prior mean of the number of substitution categories is 2 (i.e. \( E(K) = 2 \)). (a) Drosophila adh; (b) vertebrate \( \beta \)-globin; (c) Leviridae coat; (d) Japanese encephalitis env; (e) flavivirus; (f) influenza; (g) HIV pol and (h) Leviridae replicase.

Figure 8. The mean partitions of the exchangeability parameters for each of the analyses. For each alignment, three mean partitions are shown, with the top most having a prior mean of 2 and the bottom most having a prior mean of 10.
of the models in the candidate pool of models may be particularly appropriate.

Unlike the fixed amino acid models, the centred model allows for the possibility that the data in hand do not conform to any particular amino acid model. The treatment of amino acid substitutions under either the centred models or the GTR model (a model ‘centred’ on the Poisson model rates) does not entail a significant computational burden when posterior probabilities are approximated using MCMC. Maximum-likelihood estimation of the substitution parameters, by contrast, is expected to be difficult not only owing to the large number of parameters to estimate but also because the likelihood surface is likely to be flat as there is little information in a small alignment about many of the exchangeability parameters. Furthermore, by using a substitution model centred about a fixed amino acid model, one can use prior information about amino acid substitution processes combined with the data at hand to produce valid phylogenetic inferences. Indeed, an alternative to producing fixed amino acid models is to produce distributions of substitution rates from databases of alignments, such as the Pandit database (Whelan et al. 2003, 2006).

Perhaps the most intriguing possibility for the analysis of amino acid models is directly motivated by work on nucleotide models that are all four-state time-reversible continuous-time Markov chains. Several different substitution models have been described, such as those described by Jukes & Cantor (1969), Kimura (1980) and Tamura & Nei (1993), which are all simply special cases of the GTR model of DNA substitution (Tavaré 1986) but with restrictions on the substitution rates. There are a total of 203 possible models of nucleotide substitution, with about a half dozen being formally described. As we have shown here, this approach can be directly extended to the analysis of amino acid data, with the total number of restrictions on substitution rates now being $6.59 \times 10^{258}$. Substitution models in this framework are considered partitions, and in our implementation, we placed a Dirichlet process prior probability distribution on the model partitions, using MCMC to explore the space of models. The number of parameters to be estimated is determined by how the amino acid substitution rates are partitioned into equivalence classes. The data suggest a relatively small number of partitions. Thus, a uniform prior on partitions, as described by Huelsenbeck et al. (2004), places too much weight on an intermediate number of rate classes and is inefficient. The Dirichlet process prior provides a flexible class of priors, where the number of classes can be controlled by a single parameter $\beta$ and the influence of the choice of $\beta$ on the posterior can be assessed. Importantly, we find that the mean number of partitions to be very similar for a variety of choices of $\beta$.

J.P.H. was supported by NIH grant GM-069801 and NSF grant DEB-0445453. P.J. was supported by NIH grant GM-076040 and NSF grant DEB-0515738.

REFERENCES


Cambridge, UK: Cambridge University Press.


London, UK: Chapman and Hall.


Phil. Trans. R. Soc. B (2008)


