PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2008) **363**, 3941–3953
doi:10.1098/rstb.2008.0175
Published online 7 October 2008

Bayesian analysis of amino acid substitution models

John P. Huelsenbeck[1,*], **Paul Joyce**[2], **Clemens Lakner**[3] **and Fredrik Ronquist**[4]

[1]*Department of Integrative Biology, University of California, Berkeley, 3060 VLSB #3140,*
Berkeley, CA 94720-3140, USA
[2]*Departments of Mathematics and Statistics, University of Idaho, Moscow, ID 83844, USA*
[3]*School of Computational Science, DSL 150-J, Florida State University, Tallahassee,*
FL 32306-4120, USA
[4]*Swedish 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 fixed amino acid models are specific to a gene or taxonomic group (e.g. the Mtmam 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 20×20 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 $20\times19=380$ 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 $20+190=210$ parameters.[1] 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), Rtrev (Dimmic *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

* Author for correspondence (johnh@berkeley.edu).

One contribution of 17 to a Discussion Meeting Issue 'Statistical and computational challenges in molecular phylogenetics and evolution'.

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

$$\mathbb{P}(\text{Parameter}|\text{Data}) = \frac{\mathbb{P}(\text{Data}|\text{Parameter})\mathbb{P}(\text{Parameter})}{\mathbb{P}(\text{Data})},$$

where $\mathbb{P}$(Parameter|Data) is the posterior probability distribution of the parameter; $\mathbb{P}$(Data|Parameter) is the likelihood; $\mathbb{P}$(Parameter) is the prior probability distribution of the parameter; and $\mathbb{P}$(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 $\boldsymbol{X}$. (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 γ (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

$$\boldsymbol{Q} = \{q_{ij}\}$$

$$= \begin{pmatrix} - & \theta_{\mathrm{AR}}\pi_{\mathrm{R}} & \theta_{\mathrm{AN}}\pi_{\mathrm{N}} & \cdots & \theta_{\mathrm{AW}}\pi_{\mathrm{W}} & \theta_{\mathrm{AY}}\pi_{\mathrm{Y}} & \theta_{\mathrm{AV}}\pi_{\mathrm{V}} \\ \theta_{\mathrm{AR}}\pi_{\mathrm{A}} & - & \theta_{\mathrm{RN}}\pi_{\mathrm{N}} & \cdots & \theta_{\mathrm{RW}}\pi_{\mathrm{W}} & \theta_{\mathrm{RY}}\pi_{\mathrm{Y}} & \theta_{\mathrm{RV}}\pi_{\mathrm{V}} \\ \theta_{\mathrm{AN}}\pi_{\mathrm{A}} & \theta_{\mathrm{RN}}\pi_{\mathrm{R}} & - & \cdots & \theta_{\mathrm{NW}}\pi_{\mathrm{W}} & \theta_{\mathrm{NY}}\pi_{\mathrm{Y}} & \theta_{\mathrm{NV}}\pi_{\mathrm{V}} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \theta_{\mathrm{AW}}\pi_{\mathrm{A}} & \theta_{\mathrm{RW}}\pi_{\mathrm{R}} & \theta_{\mathrm{NW}}\pi_{\mathrm{N}} & \cdots & - & \theta_{\mathrm{WY}}\pi_{\mathrm{Y}} & \theta_{\mathrm{WV}}\pi_{\mathrm{V}} \\ \theta_{\mathrm{AY}}\pi_{\mathrm{A}} & \theta_{\mathrm{RY}}\pi_{\mathrm{R}} & \theta_{\mathrm{NY}}\pi_{\mathrm{N}} & \cdots & \theta_{\mathrm{YW}}\pi_{\mathrm{W}} & - & \theta_{\mathrm{YV}}\pi_{\mathrm{V}} \\ \theta_{\mathrm{AV}}\pi_{\mathrm{A}} & \theta_{\mathrm{RV}}\pi_{\mathrm{R}} & \theta_{\mathrm{NV}}\pi_{\mathrm{N}} & \cdots & \theta_{\mathrm{WV}}\pi_{\mathrm{W}} & \theta_{\mathrm{YV}}\pi_{\mathrm{Y}} & - \end{pmatrix} \mu,$$

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: $\boldsymbol{S}=$(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 θ_{AR}, θ_{AN}, θ_{AD}, …, θ_{YV}.

The 20 amino acid frequencies, $\pi=(\pi_{\mathrm{A}}, \pi_{\mathrm{R}}, \pi_{\mathrm{N}}, \ldots, \pi_{\mathrm{V}})$ are, of course, constrained to sum to one. The 190 exchangeability parameters, $\theta=(\theta_{\mathrm{AR}}, \theta_{\mathrm{AN}}, \theta_{\mathrm{AD}}, \ldots, \theta_{\mathrm{YV}})$, 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 $\theta_{\mathrm{AR}}=1$, $\theta_{\mathrm{AN}}=3$, $\theta_{\mathrm{AD}}=2$, … are equivalent to $\theta_{\mathrm{AR}}=2$, $\theta_{\mathrm{AN}}=6$, $\theta_{\mathrm{AD}}=4$, … In this study, we impose the constraint that the 190 rate parameters sum to one. The rate matrix is scaled by a factor

μ to ensure that the mean rate of substitution is one. Branch lengths of the tree, then, are interpreted as the expected number of amino acid substitutions per site. The scaling factor is equal to $\mu = -1/\sum_{\boldsymbol{S}} \pi_s q_{ss}$, where $\boldsymbol{S} \in (\mathrm{A}, \mathrm{R}, \mathrm{N}, \ldots, \mathrm{W}, \mathrm{Y}, \mathrm{V})$.

We assume that both the amino acid frequency parameters, π, and the exchangeability parameters, θ, follow different Dirichlet prior probability distributions. The use of a Dirichlet prior probability distribution for the exchangeability parameters of the GTR model of DNA substitution was first described by Zwickl & Holder (2004). Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_K)$ be K random variables that are constrained to sum to one. The Dirichlet probability density is

$$f(\boldsymbol{x}|\boldsymbol{\alpha}) = \frac{1}{b(\boldsymbol{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1},$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)$ are the parameters of the distribution and the constant of integration is a well-known ratio of gamma functions given by

$$b(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\alpha_0)}.$$

($\alpha_0 = \sum_{i=1}^{K} \alpha_i$).The marginal probability distribution of the ith Dirichlet random variable is a beta distribution with parameters α_i and $\alpha_0 - \alpha_i$. The expected value for the ith Dirichlet random variable is $E(X_i) = \alpha_i/\alpha_0$ and the variance is $\mathrm{var}(X_i) = (\alpha_i(\alpha_0 - \alpha_i)/\alpha_0^2(\alpha_0 + 1))$.

We assume that the amino acid frequencies follow a flat Dirichlet prior distribution in which $\alpha_1 = \alpha_2 = \cdots = \alpha_{20} = 1$ ($K = 20$). We consider several different prior distributions for the exchangeability parameters, $\boldsymbol{\theta}$. In general, the exchangeability parameters have the following Dirichlet probability distribution:

$$f(\boldsymbol{\theta}|\chi\boldsymbol{\nu}) = \frac{1}{b(\chi\boldsymbol{\nu})} \prod_{i<j \in \boldsymbol{S}} \theta_{ij}^{\chi\nu_{ij} - 1},$$

where $\boldsymbol{S}$ is the set of ordered amino acid states. The constant of integration is, again, a ratio of gamma functions. The parameters ν_{ij} are constrained to sum to one. We interpret the parameter χ as a concentration parameter that controls the variance of the Dirichlet prior distribution. The expectation and variance of the exchangeability parameters are $E(\theta_{ij}) = \nu_{ij}$ and $\mathrm{var}(\theta_{ij}) = (\nu_{ij}(1 - \nu_{ij})/\chi + 1)$, respectively. Note that the maximum variance allowed is $\nu_{ij}(1 - \nu_{ij})$ that occurs at the improper prior $\chi = 0$ and is the variance associated with a single draw from a binomial with proportion ν_{ij}. One can also show that $\mathrm{cov}(\theta_{ij}, \theta_{mn}) = -(\nu_{ij}\nu_{mn}/\chi + 1)$; θ_{ij} and θ_{mn} will be most correlated when $\chi = 0$ which corresponds to the covariance associated with a single draw from a multinomial $(\nu_{ij}, \nu_{mn}, 1 - \nu_{ij} - \nu_{mn})$.

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 ν_{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

$$\boldsymbol{Q} = \begin{pmatrix} - & 1.0\pi_1 & 1.2\pi_2 \\ 1.0\pi_0 & - & 1.4\pi_2 \\ 1.2\pi_0 & 1.4\pi_1 & - \end{pmatrix} \mu.$$

A Dirichlet prior distribution centred on this rate matrix would have $\nu_{01} = 1.0/(1.0 + 1.2 + 1.4) = 0.278$, $\nu_{02} = 1.2/(1.0 + 1.2 + 1.4) = 0.333$ and $\nu_{12} = 1.4/(1.0 + 1.2 + 1.4) = 0.389$. The concentration parameter, χ, specifies how the probability density is spread around the centred rate value. When χ 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 χ put very little prior probability density on rates that are different from those specified by the fixed model. Note, in fact, that as $\chi \to \infty$, 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

$$\binom{20}{2} = 190,$$

centred rates instead of

$$\binom{3}{2} = 3,$$

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 $\nu_{ij} = 1/190$ and $\chi = 190$. Similarly, we can explore the sensitivity of rate parameter estimates centred on fixed amino acid models by varying the concentration parameter χ.

(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), Mtmam (M_5; Cao *et al.* 1998; Yang *et al.* 1998), WAG (M_6; Whelan & Goldman 2001), Rtrev (M_7; Dimmic *et al.* 2002), Cprev (M_8; Adachi *et al.* 2000), Vt (M_9; Muller & Vingron 2000) and Blosum (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

$$f(\tau_i, v_i, M_j|\boldsymbol{X}) = \frac{f(\boldsymbol{X}|\tau_i, \boldsymbol{v}_i, M_j) f(\boldsymbol{v}_i) \frac{1}{T(s)} \frac{1}{10}}{\sum_{j=1}^{10} \left(\sum_{j=1}^{10} \left(\int_{v_i} f(\boldsymbol{X}|\tau_i, \boldsymbol{v}_i, M_j) f(\boldsymbol{v}_i) \right) \mathrm{d}\boldsymbol{v}_i \frac{1}{T(s)} \right) \frac{1}{10}},$$

where $f(\boldsymbol{X}|\tau_i, \boldsymbol{v}_i, M_j)$ is the likelihood for the ith tree and jth amino acid model; $f(v_i)$ 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, *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 ith tree is obtained by integrating over the other model parameters:

$$f(\tau_i|\boldsymbol{X}) = \sum_{j=1}^{10} \left(\int_{\boldsymbol{v}_i} f(\tau_i, \boldsymbol{v}_i, M_j|\boldsymbol{X}) \, \mathrm{d}\boldsymbol{v}_i \right).$$

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.

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 analysis of amino acid data; in addition to proposal mechanisms that change the tree and branch lengths, we also include a proposal mechanism that changes the amino acid rate matrix that is used to calculate likelihoods. The proposal mechanism works as follows. The current amino acid model is denoted M. We propose a new model, denoted M', by choosing one of the other nine models with equal probability. The proposed model is accepted with probability

$$\begin{aligned} R &= \min\left(1, \frac{f(\boldsymbol{X}|\tau_i, \boldsymbol{v}_i, M')}{f(\boldsymbol{X}|\tau_i, \boldsymbol{v}_i, M)} \times \frac{f(M')}{f(M)} \times \frac{f(M|M')}{f(M'|M)}\right) \\ &= \min\left(1, \frac{f(\boldsymbol{X}|\tau_i, \boldsymbol{v}_i, M')}{f(\boldsymbol{X}|\tau_i, \boldsymbol{v}_i, M)} \times \frac{1/10}{1/10} \times \frac{1/9}{1/9}\right) \\ &= \min\left(1, \frac{f(\boldsymbol{X}|\tau_i, \boldsymbol{v}_i, M')}{f(\boldsymbol{X}|\tau_i, \boldsymbol{v}_i, M)}\right). \end{aligned}$$

In a Bayesian analysis, model choice is often guided by Bayes factors. The Bayes factor for a comparison of two models (M_1 and M_2) is calculated as the ratio of the marginal likelihoods

$$BF_{12} = \frac{f(\boldsymbol{X}|M_1)}{f(\boldsymbol{X}|M_2)} = \frac{\frac{f(M_1|\boldsymbol{X})}{f(M_2|\boldsymbol{X})}}{\frac{f(M_1)}{f(M_2)}}.$$

The Bayes factor measures the 'the *change* in the odds in favour of the hypothesis when going from the prior to the posterior' (Lavine & Schervish 1999, p. 120). In this study, we calculate the Bayes factor for each model, against all of the other models. The Bayes factor for amino acid model i, then, is

$$BF_i = \frac{\frac{f(M_i|\boldsymbol{X})}{1-f(M_i|\boldsymbol{X})}}{\frac{f(M_i)}{1-f(M_i)}}.$$

(c) *Considering amino acid models as partitions*

Consider, again, the three-state model of change with instantaneous rate matrix

$$\boldsymbol{Q} = \begin{pmatrix} - & \theta_{01}\pi_1 & \theta_{02}\pi_2 \\ \theta_{01}\pi_0 & - & \theta_{12}\pi_2 \\ \theta_{02}\pi_0 & \theta_{12}\pi_1 & - \end{pmatrix}\mu$$

and exchangeability parameters θ_{ij}, where $0 \le i \le j \le 2$. This rate matrix is analogous to the GTR model of DNA or amino acid substitution, but with three states instead of 4 or 20. The exchangeability parameters can be restricted to give submodels of the most general model. For example, if rates are constrained to be equal, with $\theta_{01}=\theta_{02}=\theta_{12}$, then the model is isomorphic to the model first described by Felsenstein (1981). Other possible models apply the restriction $\theta_{01}=\theta_{02}$, $\theta_{01}=\theta_{12}$ 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

01	02	12
1	1	1
1	1	2
1	2	1
1	2	2
1	2	3

.

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 θ_{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

$$\mathcal{B}(n) = \sum_{k=0}^{n} \mathcal{S}_2(n,k),$$

where $\mathcal{S}_2(n,k)$ is the Stirling number of the second kind

$$\mathcal{S}_2(n,k) = \frac{1}{k!}\sum_{i=0}^{k-1}(-1)^i\binom{k}{i}(k-i)^n$$

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 $\mathcal{B}(3)=5$. Similarly, when only the exchangeability parameters are considered, there are a total of $\mathcal{B}(6)=203$ 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 $\mathcal{B}(190)=6.59\times 10^{258}$ 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 6.59×10^{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 with $k=48$ substitution types have a probability of $\mathcal{S}_2(190,48)/\mathcal{B}(190)=0.133$, whereas models with only $k=3$ rate classes have a probability of $\mathcal{S}_2(190,3)/\mathcal{B}(190)=1.14\times10^{-169}$. 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 probability 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 probability that a new rate class is added at the ith step be $\beta/(i-1+\beta)$. If we keep track of the number of rate classes added, by letting Y_i be a binary random variable that is equal to one, when a new rate class is added, then the total number of rate classes is $K=\sum_{i=1}^{n} Y_i$. Here n is the total number of possible rates ($n=190$). Therefore, $E(K)=\beta\sum_{i=1}^{n}\frac{1}{i-1+\beta}\approx\beta\ln(i+\beta)$. Small values for β result in only a few groups of substitution types (k) whereas large values of β 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) β-globin sequences from $s=17$ vertebrates (Yang *et al.* 2000), (iii) coat protein sequences from $s=9$ viruses from the family Leviviridae (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) $s=28$ 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 Leviviridae (Bollback & Huelsenbeck 2001) and (viii) E-glycoprotein sequences sampled from $s=18$ flavivirus (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 $\left(\sqrt{\hat{R}}\right)$ 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.

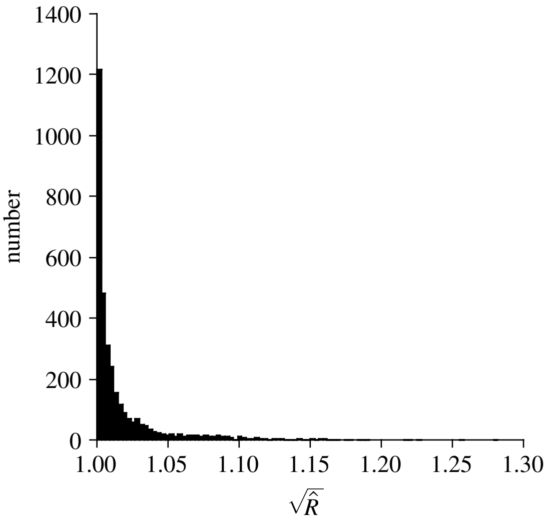

Figure 1. A frequency histogram of the potential scale reduction statistics $\left(\sqrt{\hat{R}}\right)$ for the parameters examined in this study.

(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 (θ), amino acid frequency parameters (π), gamma-shape parameter (γ), phylogenetic tree (τ) and branch-length parameters (v).

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

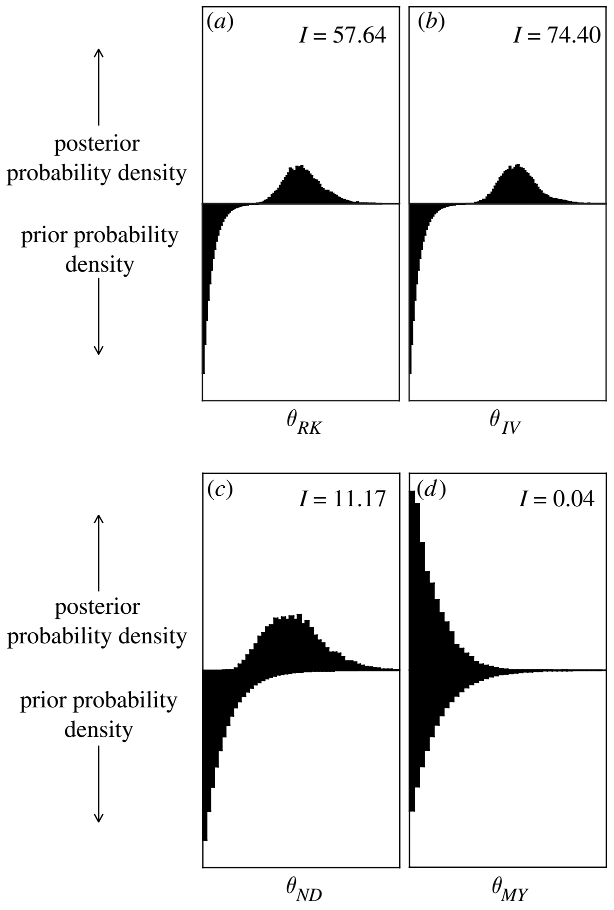

Figure 2. The marginal prior and posterior probability distribution for four of the exchangeability parameters for the HIV *pol* alignment. The Kullback–Leibler divergence, *I*, is shown for each parameter and measures the dissimilarity between the prior and posterior probability distributions. Note that the Kullback–Leibler divergence is small for the $M \leftrightarrow Y$ parameter where the prior and posterior distributions are similar. The data are more informative for the other parameters shown, and the Kullback–Leibler divergence is correspondingly larger.

prior probability distribution; there is very little information in the data about this particular parameter. However, for the other three exchangeability parameters shown in figure 2, the data are informative and the marginal posterior probability distribution is shifted away from the prior distribution.

Summarizing the results of a Bayesian analysis of a parameter-rich model can be difficult. We summarize the results of the Bayesian analyses of the exchangeability parameters in two ways. First, we examine the mean of the marginal posterior probability distribution for each rate. Second, we also calculate the Kullback–Leibler divergence (Kullback & Leibler 1951) between the prior and posterior probability distributions for each exchangeability parameter. The Kullback–Leibler divergence between two continuous probability distributions, $f(x)$ and $g(x)$, is defined as

$$I(f,g) = \int f(x) \ln\left(\frac{f(x)}{g(x)}\right) \mathrm{d}x,$$

where integration is over all possible values of the random variable x. We assume that the posterior probability distribution can be closely approximated by a beta distribution with parameters that are 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

$$I = \ln\frac{b(a_2,b_2)}{b(a_1,b_1)} - (a_2 - a_1)\psi(a_1) - (b_2 - b_1)\psi(b_1) + (a_2 - a_1 + b_2 - b_1)\psi(a_1 + b_1),$$

where, as earlier, the beta function $b(x,y) = \Gamma(x)\Gamma(y)/\Gamma(x+y)$ is the ratio of gamma functions and $\psi(x) = \Gamma'(x)/\Gamma(x)$ is the digamma function. Figure 2 shows the Kullback–Leibler divergence between the prior and posterior distributions for the exchangeability parameters, $R \leftrightarrow K$, $I \leftrightarrow V$, $N \leftrightarrow D$ and $M \leftrightarrow Y$. Note that the Kullback–Leibler divergence is large when the prior and posterior distributions are dissimilar. 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 are 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), Blosum (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, χ, 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 $I \leftrightarrow V$ (θ_{IV}) and $M \leftrightarrow F$ (θ_{MF}) parameters for the HIV alignment. In each case, the exchangeability parameters were chosen to display an instance in which the data are informative ($I \leftrightarrow V$) and rather uninformative ($M \leftrightarrow F$) 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

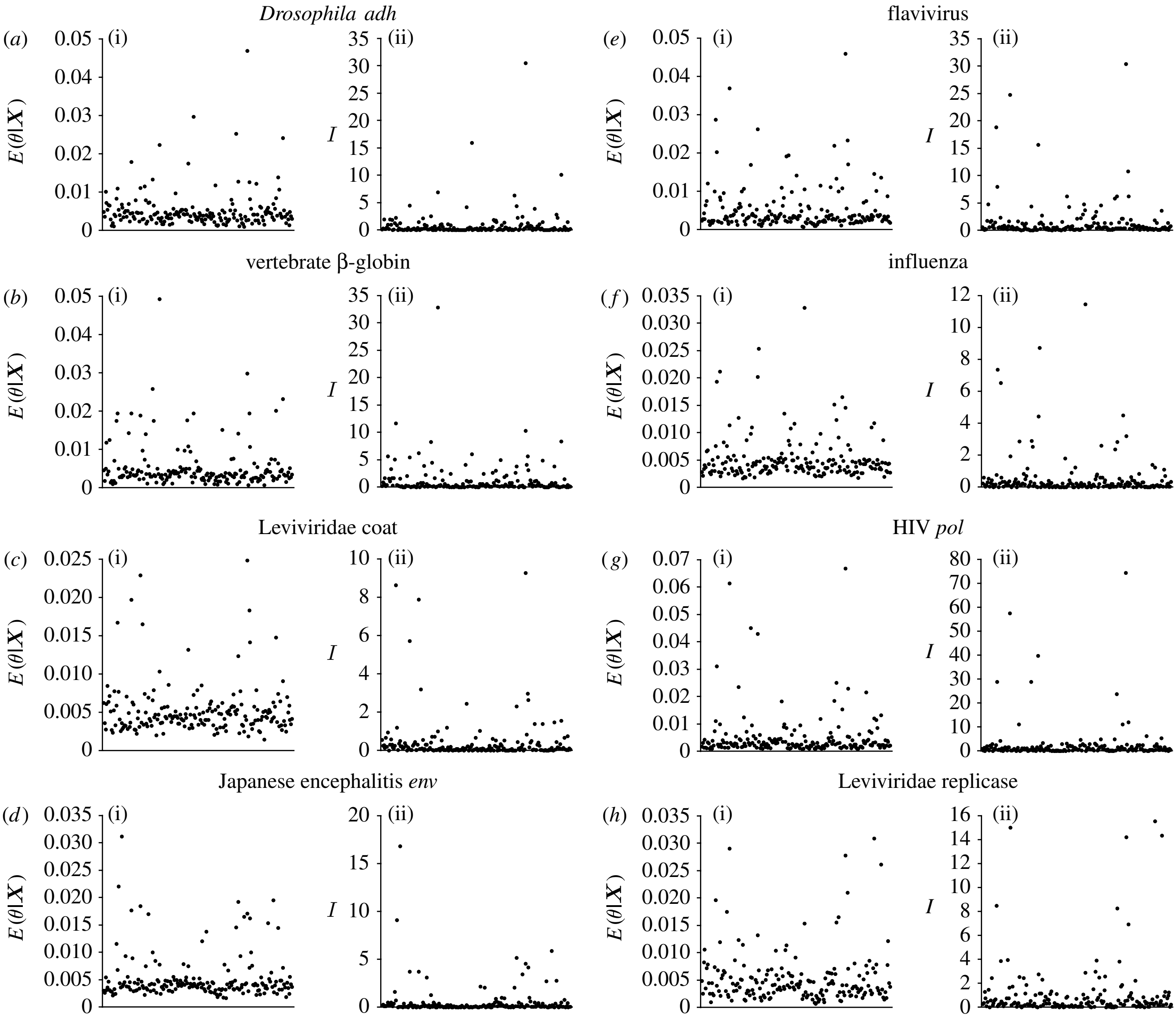

Figure 3. The mean of the marginal posterior probability distribution [[$E(\theta|\boldsymbol{X})$]] and the Kullback–Leibler divergence (I) for the 190 exchangeability parameters for each of the eight alignments examined in this study. The 190 exchangeability parameters are ordered along the x-axes.

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, χ, is increased.

The concentration parameter, χ, strongly influences the posterior probability distributions of the exchangeability parameters. Figure 6 shows the marginal posterior probability distribution for one of the exchangeability parameters ($I \leftrightarrow V$) for the HIV alignment when the prior is centred on the Poisson model. The marginal posterior distributions are most similar when χ is small (e.g. when $\chi = 0.1$ and $\chi = 1$), but quite different for larger values of χ. 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 $\alpha_1 = 1$, $\alpha_2 = 1$, $\alpha_3 = 1$ and $\alpha_4 = 1$ essentially assumes four prior observations (one for each of the four categories). In our use of the Dirichlet prior distribution, the χ parameter can be interpreted as the prior number of observations. Hence, a value of $\chi = 100$ might be interpreted as there being 100 prior observations. For many datasets, even a value of $\chi = 100$ might be considered large, and potentially swamp the information in the alignment about the values of the exchangeability parameters. Small values of χ, 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

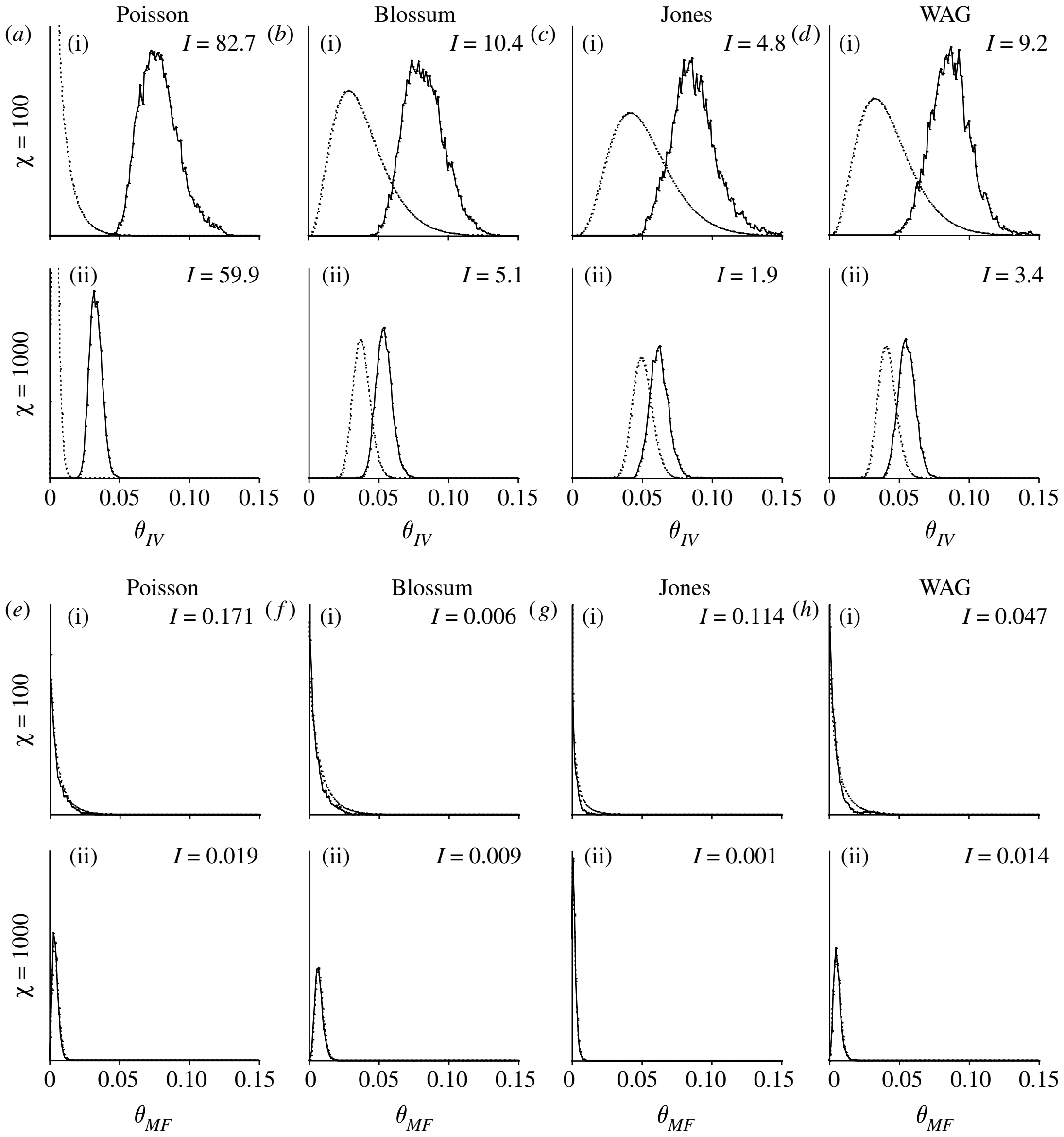

Figure 4. The prior (dashed line) and posterior (solid line) marginal probability density for two of the 190 exchangeability parameters for analyses of the HIV alignment. The *y*-axis is the marginal posterior probability density of the rate.

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 6.59×10^{258} possible time-reversible amino acid substitution models and explored the consequence of using different values for the concentration parameter of the Dirichlet process prior (β). Specifically, we fixed β 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

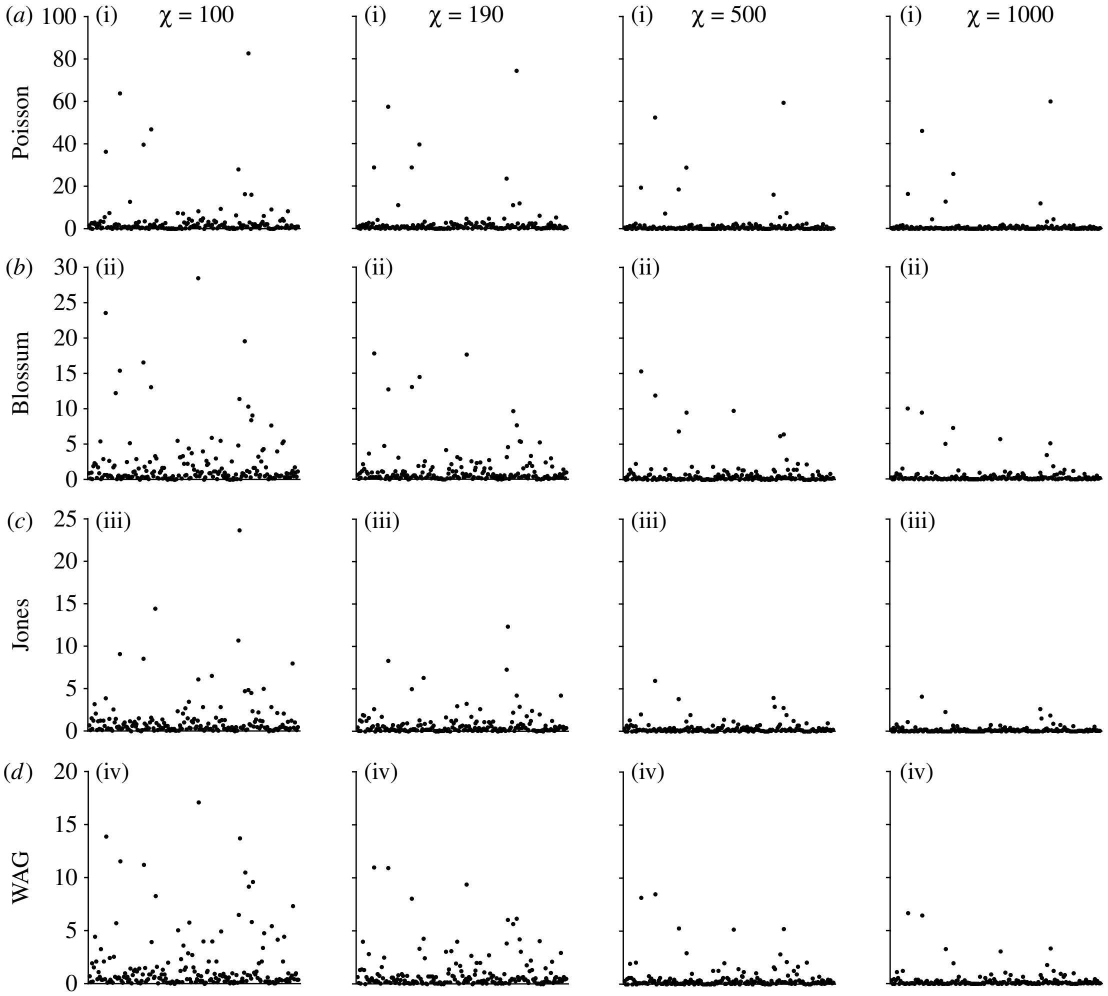

Figure 5. The Kullback–Leibler divergences (y-axes) of the 190 exchangeability parameters (arranged along the x-axes) under four different centred prior distributions for the HIV alignment.

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 β 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

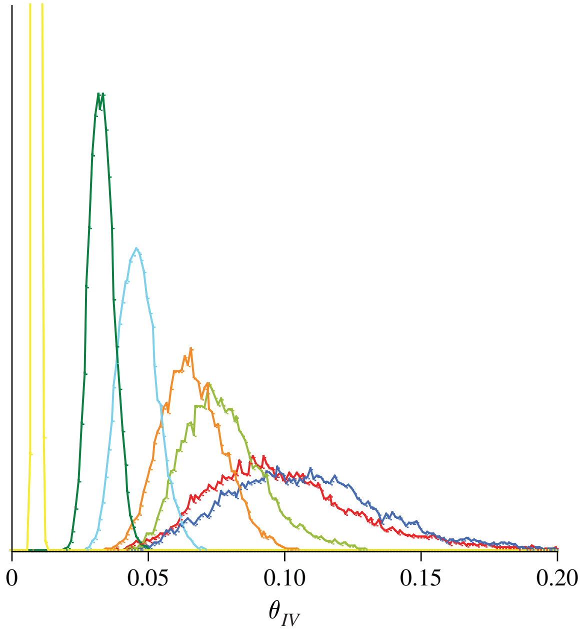

Figure 6. The marginal posterior probability distribution of the $I \leftrightarrow V$ exchangeability parameter for the HIV alignment when χ varies. The y-axis is the marginal posterior probability density of the rate. Red line, 0.1; blue line, 1; light green line, 100; orange line, 190; light blue 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, Leviviridae coat; D, Japanese encephalitis *env*; E, flavivirus; F, influenza; G, HIV *pol*; H, Leviviridae replicase.)

model	A	B	C	D	E	F	G	H
M_1	6.20×10^{-101}	3.69×10^{-119}	1.33×10^{-66}	1.52×10^{-72}	1.56×10^{-151}	9.01×10^{-55}	5.35×10^{-254}	2.54×10^{-95}
M_2	6.20×10^{-3}	1.04×10^{-10}	1.41×10^{-8}	9.99×10^{-1}	9.99×10^{-1}	9.99×10^{-1}	9.99×10^{-1}	1.69×10^{-10}
M_3	2.81×10^{-11}	9.99×10^{-1}	1.39×10^{-8}	2.11×10^{-15}	1.71×10^{-18}	1.27×10^{-14}	4.12×10^{-45}	1.09×10^{-26}
M_4	5.55×10^{-34}	2.59×10^{-53}	1.52×10^{-60}	3.40×10^{-33}	2.47×10^{-87}	3.16×10^{-33}	3.04×10^{-131}	3.95×10^{-91}
M_5	5.46×10^{-44}	8.65×10^{-78}	7.53×10^{-89}	3.72×10^{-32}	1.92×10^{-113}	5.07×10^{-37}	2.56×10^{-143}	6.65×10^{-121}
M_6	9.88×10^{-1}	5.04×10^{-6}	9.99×10^{-1}	4.10×10^{-8}	7.81×10^{-6}	1.54×10^{-8}	7.22×10^{-31}	2.27×10^{-2}
M_7	4.53×10^{-3}	1.04×10^{-13}	4.14×10^{-9}	2.68×10^{-32}	3.21×10^{-20}	4.11×10^{-24}	1.99×10^{-12}	6.91×10^{-10}
M_8	6.67×10^{-4}	5.65×10^{-19}	1.97×10^{-9}	2.99×10^{-10}	5.26×10^{-7}	1.20×10^{-8}	6.49×10^{-17}	5.68×10^{-10}
M_9	2.24×10^{-7}	1.77×10^{-27}	5.14×10^{-5}	1.34×10^{-15}	9.83×10^{-17}	2.96×10^{-17}	1.31×10^{-44}	4.28×10^{-18}
M_{10}	1.51×10^{-10}	8.13×10^{-24}	4.43×10^{-5}	6.40×10^{-23}	2.21×10^{-20}	4.00×10^{-22}	3.12×10^{-67}	9.77×10^{-1}

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, Leviviridae coat; D, Japanese encephalitis *env*; E, flavivirus; F, influenza; G, HIV *pol*; H, Leviviridae replicase.)

model	A	B	C	D	E	F	G	H
M_1	5.58×10^{-100}	3.32×10^{-118}	1.19×10^{-65}	1.41×10^{-71}	1.41×10^{-150}	8.11×10^{-54}	4.81×10^{-253}	2.29×10^{-94}
M_2	5.61×10^{-2}	9.41×10^{-10}	1.26×10^{-7}	2.17×10^{8}	1.08×10^{6}	3.26×10^{8}	4.51×10^{12}	1.52×10^{-9}
M_3	2.53×10^{-10}	1.78×10^{6}	1.25×10^{-7}	1.89×10^{-14}	1.54×10^{-17}	1.14×10^{-13}	3.71×10^{-44}	9.81×10^{-26}
M_4	5.00×10^{-33}	2.33×10^{-52}	1.37×10^{-59}	3.06×10^{-32}	2.22×10^{-86}	2.84×10^{-32}	2.73×10^{-130}	3.56×10^{-90}
M_5	4.91×10^{-43}	7.78×10^{-77}	6.78×10^{-88}	3.34×10^{-31}	1.73×10^{-112}	4.56×10^{-36}	2.31×10^{-142}	5.99×10^{-120}
M_6	7.80×10^{2}	4.53×10^{-5}	9.39×10^{4}	3.69×10^{-7}	7.03×10^{-5}	1.39×10^{-7}	6.49×10^{-30}	2.09×10^{1}
M_7	4.09×10^{-2}	9.44×10^{-13}	3.73×10^{-8}	2.41×10^{-31}	2.89×10^{-19}	3.70×10^{-23}	1.79×10^{-11}	6.22×10^{-9}
M_8	6.01×10^{-3}	5.09×10^{-18}	1.77×10^{-8}	2.69×10^{-9}	4.74×10^{-6}	1.08×10^{-7}	5.84×10^{-16}	5.12×10^{-9}
M_9	2.02×10^{-6}	1.60×10^{-26}	4.62×10^{-4}	1.21×10^{-14}	8.85×10^{-16}	2.66×10^{-16}	1.17×10^{-43}	3.85×10^{-17}
M_{10}	1.36×10^{-9}	7.32×10^{-23}	3.98×10^{-4}	5.76×10^{-22}	1.98×10^{-19}	3.60×10^{-21}	2.81×10^{-66}	3.86×10^{2}

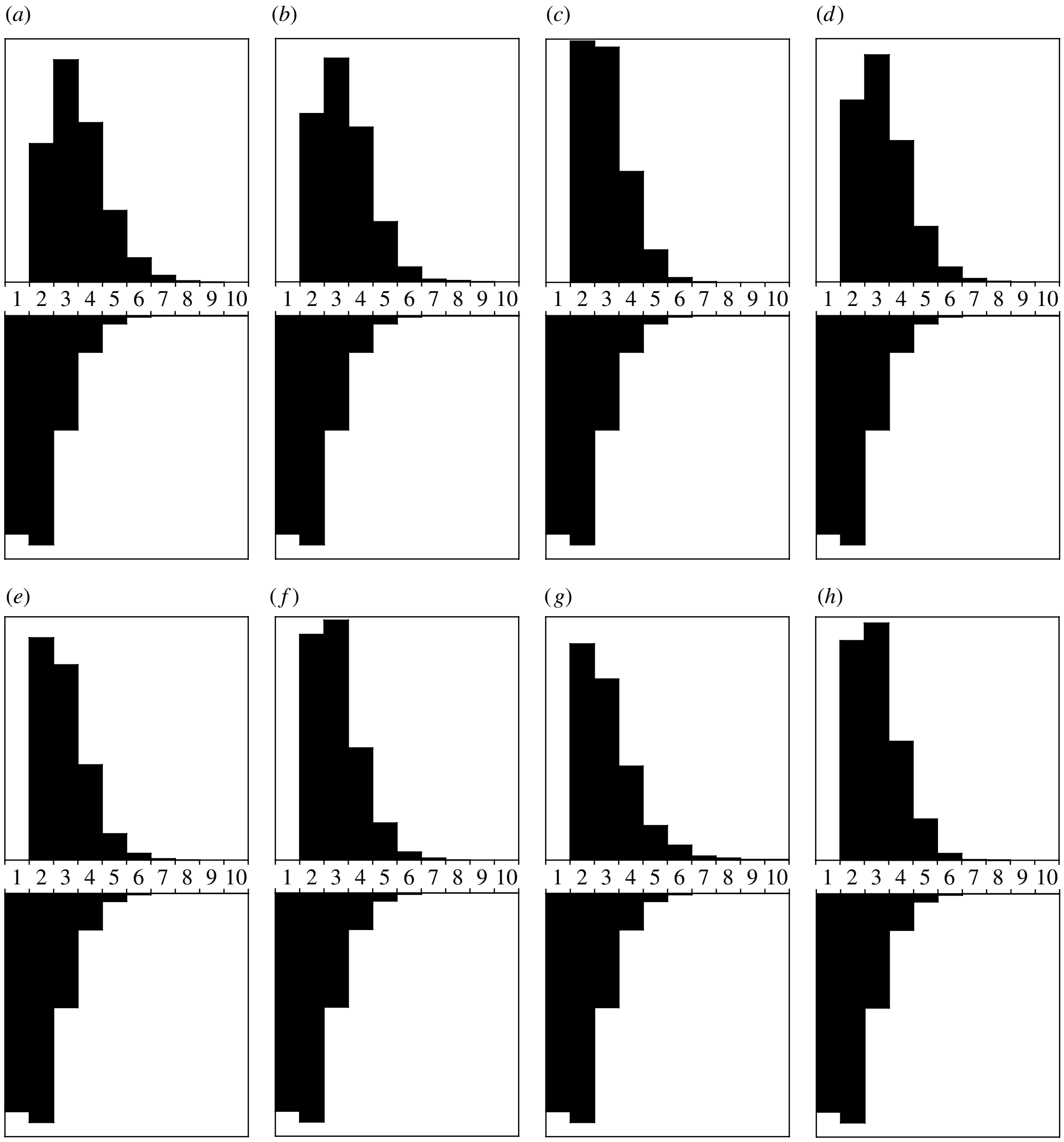

Figure 7. The prior and posterior probability distributions (upside down and right side up, respectively) for the number of exchangeability parameter groups for the analyses where the concentration parameter is fixed such that the prior mean of the number of substitution categories is 2 (i.e. $E(K)=2$).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

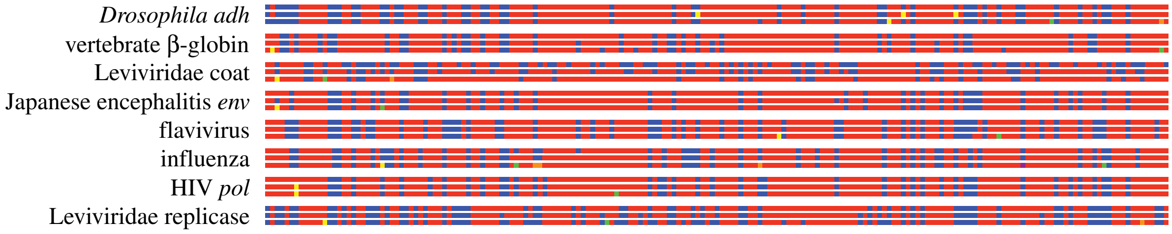

Figure 8. The mean partitions of the exchangeability parameters for each of the analyses. For each alignment, three mean partitions are shown, with the top most having a prior mean of 2 and the bottom most having a prior mean of 10.

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 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 6.59×10^{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 β and the influence of the choice of β 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 β.

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.

ENDNOTE

[1] Of the 210 parameters, 208 are free to vary. The amino acid frequency parameters are constrained to sum to one, so knowledge of 19 of them is sufficient. Most implementations of the GTR model involve rescaling the substitution rates to be one, so only the relative values of the exchangeability parameters influence the likelihood and one loses a free parameter from the list of exchangeability parameters. In many implementations, one of the substitution rates is set to one, and the others are measured relative to that value. Here, we will persist in treating all 190 exchangeability parameters as if they were independently estimated.

REFERENCES

Adachi, J. & Hasegawa, M. 1996 MOLPHY v. 2.3: programs for molecular phylogenetics based on maximum likelihood. *Comput. Sci. Monogr.* **28**, 1–150.

Adachi, J., Waddell, P., Martin, W. & Hasegawa, M. 2000 Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. *J. Mol. Evol.* **50**, 348–358.

Antoniak, C. E. 1974 Mixtures of Dirichlet processes with applications to non-parametric problems. *Ann. Stat.* **2**, 1152–1174. (doi:10.1214/aos/1176342871)

Bell, E. T. 1934 Exponential numbers. *Am. Math. Monthly* **41**, 411–419. (doi:10.2307/2300300)

Bishop, M. J. & Friday, A. E. 1987 Tetrapod relationships: the molecular evidence. In *Molecules and morphology in evolution: conflict or compromise?* (ed. C. Patterson), pp. 123–140. Cambridge, UK: Cambridge University Press.

Bollback, J. & Huelsenbeck, J. P. 2001 Phylogenetic relationships, genome evolution, and host specificity of single-stranded RNA bacteriophage (Family Leviviridae). *J. Mol. Evol.* **52**, 117–128.

Cao, Y., Janke, A., Waddell, P. J., Westerman, M., Takenaka, O., Murata, S., Okada, N., Pääbo, S. & Hasegawa, M. 1998 Conflict amongst individual mitochondrial proteins in resolving the phylogeny of eutherian orders. *J. Mol. Evol.* **47**, 307–322. (doi:10.1007/PL00006389)

Cornish-Bowden, A. 1985 Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res.* **13**, 3021–3030. (doi:10.1093/nar/13.9.3021)

Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. 1978 A model of evolutionary change in proteins. In: *Atlas of protein sequence and structure*, vol. 5, Supplement 3, pp. 345–352. Washington, DC: National Biomedical Research Foundation.

Dimmic, M. W., Rest, J. S., Mindell, D. P. & Goldstein, D. 2002 RArtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. *J. Mol. Evol.* **55**, 65–73. (doi:10.1007/s00239-001-2304-y)

Felsenstein, J. 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376. (doi:10.1007/BF01734359)

Ferguson, T. S. 1973 A Bayesian analysis of some nonparametric problems. *Ann. Stat.* **1**, 209–230. (doi:10.1214/aos/1176342360)

Fitch, W. M., Bush, R. M., Bender, C. A. & Cox, N. J. 1997 Long term trends in the evolution of H(3) HA1 human influenza type A. *Proc. Natl Acad. Sci. USA* **94**, 7712–7718. (doi:10.1073/pnas.94.15.7712)

Gelman, A. 1996 Inference and monitoring convergence. In *Markov Chain Monte Carlo in practice* (eds W. R. Gilks, S. Richardson & D. J. Spiegelhalter), pp. 131–143. London, UK: Chapman and Hall.

Gelman, A. & Rubin, D. B. 1992 Inferences from iterative simulation using multiple sequences. *Stat. Sci.* **7**, 457–511. (doi:10.1214/ss/1177011136)

Gusfield, D. 2002 Partition-distance: a problem and class of perfect graphs arising in clustering. *Inf. Process. Lett.* **82**, 159–164. (doi:10.1016/S0020-0190(01)00263-0)

Hastings, W. K. 1970 Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109. (doi:10.1093/biomet/57.1.97)

Henikoff, S. & Henikoff, J. G. 1992 Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA* **89**, 10 915–10 919. (doi:10.1073/pnas.89.22.10915)

Huelsenbeck, J. P. & Andolfatto, P. 2007 Inference of population structure under a Dirichlet process model. *Genetics* **175**, 1787–1802. (doi:10.1534/genetics.106.061317)

Huelsenbeck, J. P., Ronquist, F., Nielsen, R. & Bollback, J. 2001 Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* **294**, 2310–2314. (doi:10.1126/science.1065889)

Huelsenbeck, J. P., Larget, B. & Alfaro, M. E. 2004 Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. *Mol. Biol. Evol.* **21**, 1123–1133. (doi:10.1093/molbev/msh123)

Huelsenbeck, J. P., Jain, S., Frost, S. W. D. & Pond, S. L. K. 2006 A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. *Proc. Natl Acad. Sci. USA* **103**, 6263–6268. (doi:10.1073/pnas.0508279103)

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

Jukes, T. H. & Cantor, C. R. 1969 Evolution of protein molecules. In *Mammalian protein metabolism* (ed. H. N. Munro), pp. 21–123. New York, NY: Academic Press.

Kimura, M. 1980 A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111–120. (doi:10.1007/BF01731581)

Kosiol, C., Holmes, I. & Goldman, N. 2007 An empirical codon model for protein sequence evolution. *Mol. Biol. Evol.* **24**, 1464–1479. (doi:10.1093/molbev/msm064)

Kullback, S. & Leibler, R. A. 1951 On information and sufficiency. *Ann. Math. Stat.* **22**, 79–86. (doi:10.1214/aoms/1177729694)

Larget, B. & Simon, D. L. 1999 Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.* **16**, 750–759.

Lartillot, N. & Philippe, H. 2004 A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **21**, 1095–1109. (doi:10.1093/molbev/msh112)

Lavine, M. & Schervish, M. J. 1999 Bayes factors: what they are and what they are not. *Am. Stat.* **53**, 119–122. (doi:10.2307/2685729)

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. 1953 Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092. (doi:10.1063/1.1699114)

Muller, T. & Vingron, M. 2000 Modeling amino acid replacement. *J. Comput. Biol.* **7**, 761–776. (doi:10.1089/10665270050514918)

Neal, R. M. 2000 Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Stat.* **9**, 249–265. (doi:10.2307/1390653)

Rambaut, A. & Drummond, A. J. 2007 Tracer v. 1.4. See http://beast.bio.ed.ac.uk/Tracer.

Ronquist, F. & Huelsenbeck, J. P. 2003 MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574. (doi:10.1093/bioinformatics/btg180)

Smith, B. J. 2007 boa: an R package for MCMC output convergence assessment and posterior inference. *J. Stat. Softw.* **21**, 1–37.

Stanton, D. & White, D. 1986 *Constructive combinatorics*. New York, NY: Springer.

Swofford, D. L. 1998 *PAUP*: phylogenetic analysis using parsimony (*and other methods)*, v. 4.0.b10. Sunderland, MA: Sinauer Associates, Inc.

Tamura, K. & Nei, M. 1993 Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**, 512–526.

Tavaré, S. 1986 Some probabilistic and statistical problems on the analysis of DNA sequences. *Lect. Math. Life Sci.* **17**, 57–86.

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

Whelan, S., de Bakker, P. & Goldman, N. 2003 Pandit: a database of protein and associated nucleotide domains with inferred trees. *Bioinformatics* **19**, 1556–1563. (doi:10.1093/bioinformatics/btg188)

Whelan, S., de Bakker, P., Quevillon, E., Rodriguez, N. & Goldman, N. 2006 PANDIT: an evolution-centric database of protein and associated nucleotide domains with inferred trees. *Nucleic Acids Res.* **34**, D327–D331. (doi:10.1093/nar/gkj087)

Yang, Z. 1993 Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* **10**, 1396–1401.

Yang, Z. 1994 Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39**, 306–314. (doi:10.1007/BF00160154)

Yang, Z., Nielsen, R. & Hasegawa, M. 1998 Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol. Biol. Evol.* **15**, 1600–1611.

Yang, Z., Nielsen, R., Goldman, N. & Pedersen, A. M. K. 2000 Codon-substitution models for heterogeneous selection pressure. *Genetics* **155**, 431–449.

Zanotto, P. M., Gould, E. A., Gao, G. F., Harvey, P. H. & Holmes, E. C. 1996 Population dynamics of flaviviruses revealed by molecular phylogenies. *Proc. Natl Acad. Sci. USA* **93**, 548–553. (doi:10.1073/pnas.93.2.548)

Zwickl, D. J. & Holder, M. T. 2004 Model parameterization, prior distributions, and the general time-reversible model in Bayesian phylogenetics. *Syst. Biol.* **53**, 877–888. (doi:10.1080/10635150490522584)