Syst. Biol. 53(4):571–581, 2004

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150490468675

A Phylogenetic Mixture Model for Detecting Pattern-Heterogeneity in Gene Sequence or Character-State Data

MARK PAGEL AND ANDREW MEADE

School of Animal and Microbial Sciences, University of Reading, Whiteknights, Reading RG6 6AJ, England; E-mail: m.pagel@rdg.ac.uk (M.P.)

Abstract.—We describe a general likelihood-based 'mixture model' for inferring phylogenetic trees from gene-sequence or other character-state data. The model accommodates cases in which different sites in the alignment evolve in qualitatively distinct ways, but does not require prior knowledge of these patterns or partitioning of the data. We call this qualitative variability in the pattern of evolution across sites "pattern-heterogeneity" to distinguish it from both a homogenous process of evolution and from one characterized principally by differences in rates of evolution. We present studies to show that the model correctly retrieves the signals of pattern-heterogeneity from simulated gene-sequence data, and we apply the method to protein-coding genes and to a ribosomal 12S data set. The mixture model outperforms conventional partitioning in both these data sets. We implement the mixture model such that it can simultaneously detect rate- and pattern-heterogeneity. The model simplifies to a homogeneous model or a rate-variability model as special cases, and therefore always performs at least as well as these two approaches, and often considerably improves upon them. We make the model available within a Bayesian Markov-chain Monte Carlo framework for phylogenetic inference, as an easy-to-use computer program. [Bayesian inference; MCMC; mixture model; phylogeny; rate-heterogeneity; secondary structure; sequence evolution]

The conventional likelihood-based approach to inferring phylogenetic trees from aligned gene-sequence or other data is to apply a single substitutional model to all sites. Popular variations on this approach include use of the gamma rate-heterogeneity model (Yang, 1994), or partitioning the data in which the investigator assigns a different substitutional model to different sites, later combining the information from the different models into a single overall likelihood.

The limitations of using a single homogeneous model of substitution to characterize all sites are obvious. If the data are nucleotides from a coding region, natural selection may constrain variability at some sites more than others (so-called purifying selection) and therefore sites will, minimally, exhibit different rates of evolution. This is one of the most compelling reasons for use of the gamma rate-heterogeneity model: its success in improving the likelihood of the data derives from its ability to allow some sites to be better characterized by a substitutional process that is speeded up relative to other sites.

Rate-heterogeneity models are less applicable to cases in which sites might be expected to evolve in qualitatively different ways, and not just vary in their overall rates of evolution. First, second, and third codon positions, for example, might have different substitutional patterns, independently of their tendency to evolve at different rates. A well-known case in which heterogeneity across sites in the pattern of evolution is predicted is in the stems and loops of ribosomal sequences. Stems frequently adopt canonical Watson-Crick base pairing giving the expectation that the frequency of transitional changes will greatly exceed those of transversional changes (Hillis and Dixon, 1991; Higgs, 1998; Savill et al., 2001). Although no specific prediction is made about loops, special substitutional models have been proposed to characterize ribosomal stem data (e.g., Schöniger and von Haeseler, 1994; Savill et al., 2001).

We call this qualitative variability in the pattern of evolution across sites "pattern-heterogeneity" to distinguish it from both a homogenous process of evolution and from one characterized principally by differences in rates of evolution. In addition to patterns expected within genes, codons, or secondary structures, concatenated sequence alignments may harbor large variation in both the pattern and rates of evolution across sites. Murphy et al. (2001), for example, analyzed an alignment of 22 genes comprising 16.4 kb of DNA to infer the mammalian phylogeny. Even an alignment of this size may soon seem small given the growth of what might be called genomic-phylogenetics, in which large portions of genomes are aligned across species. Rokas et al. (2003) used 106 genes comprising 127,026 sites to infer the phylogeny of the yeast.

Heterogeneity across sites in the pattern of gene-sequence evolution is often accommodated by partitioning the data such that different models of evolution are assigned to different sites. This can be helpful when there is clear evidence that the partitions follow different evolutionary models, or even necessary if qualitatively different characters, such as gene sequences and morphological traits, are combined in one analysis. In other instances, however, it may not be obvious how to partition data or it may be that there is important variability within partitions. We give examples below in which partitioning by gene, by codon position, or by the stems and loops in ribosomal data would miss significant evolutionary variation within these categories (see also Hickson et al., 1996, and Simon et al., 1994, for examples with ribosomal data).

As opposed to partitioning data, a realistic accounting of one's knowledge in many circumstances would be to entertain the possibility that different models can apply with varying probabilities to the same site in the gene or alignment. The likelihood approach is then to sum the likelihood over these different models, each weighted by its probability. The probability that a given model applies to a given site might be obtained from prior information or, as we will show, the weights can be estimated from

the data. Summing over models may be preferred when there is not a clear case for partitioning the data, and may allow for unforeseen patterns of evolution to emerge.

Gelman et al. (1995) use the term 'mixture models' to describe the practice of calculating likelihoods by summing over a range of statistical models for a given data point. In the context of phylogenetic inference, Koshi and Goldstein (1998) report a mixture model to characterize amino acid sequences and Huelsenbeck and Nielsen (1999) assume a gamma distribution of transition/transversion ratios at each site in a nucleotide sequence. Krajewski et al. (1999) attempt to construct a mixture model based upon distance matrices, and Yang et al. (2000) use a mixture model to sum over different values of the synonymous/nonsynonymous substitution ratio at each site. More recently, Lartillot and Philippe (2004) develop a mixture model for amino acid sequences that allows the equilibrium distribution of amino acid frequencies to vary among sites.

Here we describe a mixture model for detecting heterogeneity across sites in the pattern of evolution of gene-sequence data, although it is applicable in principle to any aligned character-state data capable of exhibiting more than one pattern of evolution. The method fits two or more qualitatively different models of sequence evolution to each site in a gene-sequence alignment, without specifying in advance the nature of the models, their relative probabilities, or having knowledge of which sites are best fit by which model.

THE MODEL

Define the likelihood of a model of gene sequence evolution as an amount proportional to the probability of the data given the model:

$$L(\mathbf{Q}) \propto P(\mathbf{D} \mid \mathbf{Q})$$

where $\mathbf{Q}$ is the substitution rate matrix that defines the model of evolution, and $\mathbf{D}$ will normally be an aligned set of sequence data. In the case of nucleotide data, $\mathbf{Q}$ is the familiar 4×4 matrix of transition rates among A, C, G, and T (e.g., Swofford et al., 1996). If the data consist of binary characters, then $\mathbf{Q}$ would be a 2×2 matrix, and for protein data $\mathbf{Q}$ is a 20×20 matrix representing the transition rates among all pairs of amino acids.

Given an aligned set of gene-sequence or other character-state data, the probability of the data in $\mathbf{D}$ is found as the product over sites of the individual probabilities of each site. By taking the product over sites, we are assuming that their evolution is independent. Considering that the the likelihood is calculated for a specific phylogenetic tree we can write the right hand side of the above equation as

$$P(\mathbf{D} \mid \mathbf{Q}, T) = \prod_i P(\mathbf{D}_i \mid \mathbf{Q}, T)$$

where the product is over all of the sites in the data matrix and T stands for the specific tree.

A mixture model for gene-sequence or amino acid data modifies this basic framework by including more than one model of evolution $\mathbf{Q}$. The probability of the data is now calculated by summing the likelihood at each site over all of the different $\mathbf{Q}$ matrices. Thus, defining the different matrices as $\mathbf{Q_1}, \mathbf{Q}_2, \ldots, \mathbf{Q}_J$ write the probability of the data under the mixture model as

$$P(\mathbf{D} \mid \mathbf{Q}_1, \mathbf{Q}_2, \ldots \mathbf{Q}_j, T) = \prod_i \sum_j w_j P(\mathbf{D}_i \mid \mathbf{Q}_j, T) \quad (1)$$

where the summation over J now specifies that the likelihood of the data at each site is summed over J separate rate or $\mathbf{Q}$ matrices, the summation being weighted by w's where $w_1 + w_2 + \cdots + w_j = 1.0$. The number of matrices, J, can be determined either by prior knowledge of how many different patterns are expected in the data, or empirically as we illustrate in a later section.

Equation 1 is a general statement about how to combine likelihoods from different models of evolution applied to the same data. It says that the observed data at a given site arose with probability w_j from the model implied by the rate parameters in $\mathbf{Q}_j$. One $\mathbf{Q}$ might, for example, contain parameters that conform to the nature of evolution that tends to predominate at coding positions, while another conforms to the patterns seen at silent sites. However, both are allowed to apply with some probability to each site.

Discrete-Gamma as a Mixture Model

The popular discrete gamma model (Yang, 1994) is a mixture model that is constrained to take a specific form. The gamma model supposes that rates of evolution vary across sites with probabilities that follow a gamma distribution. This is a class of right-skewed curves, reflecting the assumption that most sites evolve relatively slowly, with a smaller number evolving at higher rates.

The discretized gamma curve supplies J multipliers ranging from slow (<1) to fast (>1). The discrete-gamma model then sums the likelihood of Equation 1 over these J categories by, in turn, multiplying the elements of the single $\mathbf{Q}$ matrix by the separate γ_j scalars; the J different $\mathbf{Q}$'s of Equation 1 all become multiples of each other in the gamma model:

$$P(\mathbf{D} \mid \mathbf{Q}, \gamma, T) = \prod_i \sum_j w_j P(\mathbf{D}_i \mid \gamma_j \mathbf{Q}, T)$$

The J gamma rates are chosen to divide the continuous gamma distribution into J equally probable parts, such that $w_1 = w_2 = \ldots = w_j = 1/J$.

This statement of the discrete-gamma model emphasises its elegance—the J multipliers are obtained from a one-parameter distribution—but also its restrictions. The amount of realism that the gamma model brings to a data set depends upon whether the variability in the data is limited to differences in rates and whether these differences conform to a gamma distribution. Other

probability distributions, such as the beta, allowing left-skewed, and even U-shaped distributions of rates can easily be incorporated into the above formalism.

More generally, a mixture model allowing the **Q** matrices to adopt any configuration will always perform at least as well as the discrete-gamma (or other distribution) model, and frequently better, although the mixture model will often require more parameters. The performance of the mixture model relative to the gamma arises because the separate **Q** matrices of the general model can always be made to conform to those that would arise under the gamma model. In the limiting case when all of the data conform to a single homogeneous process, both the general mixture model and the gamma rate-heterogeneity models simplify to a model based upon a single **Q** matrix.

Combining Rate- and Pattern-Heterogeneity

To combine variation across sites in the rates of evolution with variation in the qualitative pattern of evolution, rewrite Equation 1 as

$$P(\mathbf{D} \mid \mathbf{Q}_1, \mathbf{Q}_2, \ldots \mathbf{Q}_J, \gamma, T) = \prod_i \sum_j w_j \Big/ k \sum_k P(\mathbf{D}_i \mid \gamma_k \mathbf{Q}_j, T) \qquad (2)$$

This model fits J separate rate matrices of the pattern-heterogeneity model, each of which is allowed to have K rates from the gamma rates model. If heterogeneity across sites in both the rate and patterns of evolution exist in the data, Equation 2 allows the rate-heterogeneity to be detected by the addition of a single parameter. This reduces the number of parameters in the model, freeing the remaining **Q** matrices to detect non–rate-related pattern-heterogeneity.

Partitioning and Mixture Models

Equation 1 can be used to understand the relationship of partitioning the data to mixture modeling. Partitioning data by applying different models to different sites is equivalent to setting to zero different w's of a mixture model at different sites. In some cases this partitioning might be justified on empirical grounds that it improves the likelihood of the data. In other cases, such as with secondary structure, or when different kinds of data are combined into a single analysis, the data are partitioned on the basis of an a priori expectation.

Applications of the Model to Simulated and Real Data

We implemented the pattern-heterogeneity mixture model with or without gamma rate variability in a Markov chain Monte carlo (MCMC) method for inferring phylogenetic trees, and studied its performance in several simulated and real data sets. Geyer (1992) and Gilks et al. (1996) discuss MCMC methods in general and (among others) Wilson and Balding (1998), Larget and Simon (1999), Huelsenbeck et al. (2001), and Pagel and Lutzoni (2002) discuss their use in phylogenetic inference. Our use of MCMC is one of convenience and preference. There is nothing in mixture models that requires use of MCMC, and the model could just as appropriately be implemented as a maximum-likelihood method. We use the MCMC approach here to estimate the posterior density or probability distribution of trees and parameters of the model of sequence evolution.

We use the general time reversible model (GTR) to characterize the transition rates among the four nucleotides (Swofford et al., 1996). For phylogenetic inference, this matrix is normally written as the product of a symmetric rate matrix **R**, and a diagonal matrix called Π. The **R** matrix contains the six rate parameters describing symmetrical rates of changes between pairs of nucleotides, and Π contains the four base frequencies (denoted π_i). Their product returns the matrix **Q** with up to 12 different transition rates among pairs of nucleotides:

$$Q_{GTR} = \mathbf{R}\Pi = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{bmatrix} -\sum_J q_{Aj}\pi_J & q_{AC}\pi_C & q_{AG}\pi_G & q_{AT}\pi_T \\ q_{AC}\pi_A & -\sum_J q_{Cj}\pi_J & q_{CG}\pi_G & q_{CT}\pi_T \\ q_{AG}\pi_A & q_{GC}\pi_C & -\sum_J q_{Gj}\pi_J & q_{GT}\pi_T \\ q_{AT}\pi_A & q_{CT}\pi_C & q_{GT}\pi_G & -\sum_J q_{Tj}\pi_J \end{bmatrix} \end{array}$$

The **R** matrix of the GTR model is conventionally specified by five free rate parameters, with the sixth, the G↔T transition, set to 1.0. Popular models of gene sequence evolution are simply modifications of **Q**. For example, the Jukes-Cantor model presumes that all of the transition rates and all the base frequencies are equal.

When using more than one rate matrix in our mixture model (Equation 1), we use the conventional five rate-parameter configuration for the first rate matrix, but then allow the successive matrices to have six free rate parameters. We use a common set of base frequency parameters across all rate matrices, estimated from the data, although it is straightforward to estimate these parameters separately for each matrix. In addition to the rate parameters, we estimate a weight term (Equation 1) for each rate matrix. Each additional GTR rate matrix in the mixture model therefore requires seven new parameters. Adding gamma rate-heterogeneity requires one parameter independently of the number of rate matrices.

In reporting results we denote the mixture model by the number of independent rate matrices (e.g., 2Q = two rate matrices) and we refer to the combined gamma rate-heterogeneity plus mixture model by denoting the number of rate matrices followed by Γ (e.g., $2Q + \Gamma$). We used four rate categories for all analyses with the gamma model.

Model Testing

We use Bayes factors (Gelman et al., 1995) to compare models in both the real and simulated data. The Bayes factor for model i compared to model j is the ratio of their marginal likelihoods. The marginal likelihood is the probability of the data given the model scaled by the model's prior probability, then integrated over all values of the model parameters. In a Bayesian phylogenetic setting the marginal likelihood is integrated over trees (T) and values of the rate parameters in the **Q** matrix:

$$P(\mathbf{D} \mid M) = \int_T \int_Q P(\mathbf{D} \mid \mathbf{Q}, T)\, p(\mathbf{Q})\, p(T) d\mathbf{Q}, dT$$

The term $P(\mathbf{D} \mid M)$ refers to the marginal probability of the data given some model M, where M includes the parameters of the substitutional process and the phylogenetic trees. The terms $p(\mathbf{Q})$ and $p(T)$ are the prior probabilities of the rate parameters and the tree, respectively. This integration is difficult because the parameters can vary continuously, so the marginal likelihood is approximated via the harmonic mean of the likelihoods of the data (Raftery, 1996) over a representative sample of parameter values, weighted by their prior probabilities. The sample is derived from the converged Markov chain.

Given marginal likelihoods for two different models the log-Bayes factor is defined as:

$$\log BF = -2 \log \left[\frac{p(D \mid M_i)}{p(D \mid M_j)} \right]$$

Using the logarithm of the Bayes factor, Raftery (1996; 165) suggests a rule of thumb of 2 to 5 as "positive" evidence for model i, and greater than 5 as "strong" evidence. Log-Bayes factors of zero indicate equivalence of two models, and less than 0 provide evidence for model j.

The Bayes factor test penalizes more complex models via the prior terms. Normally each prior is a number less than one. More complex models multiply together a larger number of these prior terms, reducing their marginal likelihoods relative to simpler models. In all of our MCMC runs we assigned uniform priors to trees and parameters of the models of sequence evolution, and an exponential prior to branch lengths. Using these priors, as a rule of thumb, each GTR rate matrix added to a mixture model requires an improvement in the log-likelihood of about 30 log-units to return a Bayes factor of 0.

Simulated Data

We used *SeqGen* (Rambaut and Grassly, 1997) to simulate nucleotide sequences on a randomly generated phylogenetic tree of 50 species with branch lengths randomly chosen to vary between 0 and 1.0. We produced data according to four models of evolution, always simulating 2000 independent sites: a homogeneous process model, a continuous gamma rates model (with the gamma shape parameter set to 1.0), a mixture model based upon two different rate matrices, and a combined pattern- and rate-heterogeneity mixture model also using two rate matrices. We simulated all models using GTR rate matrices.

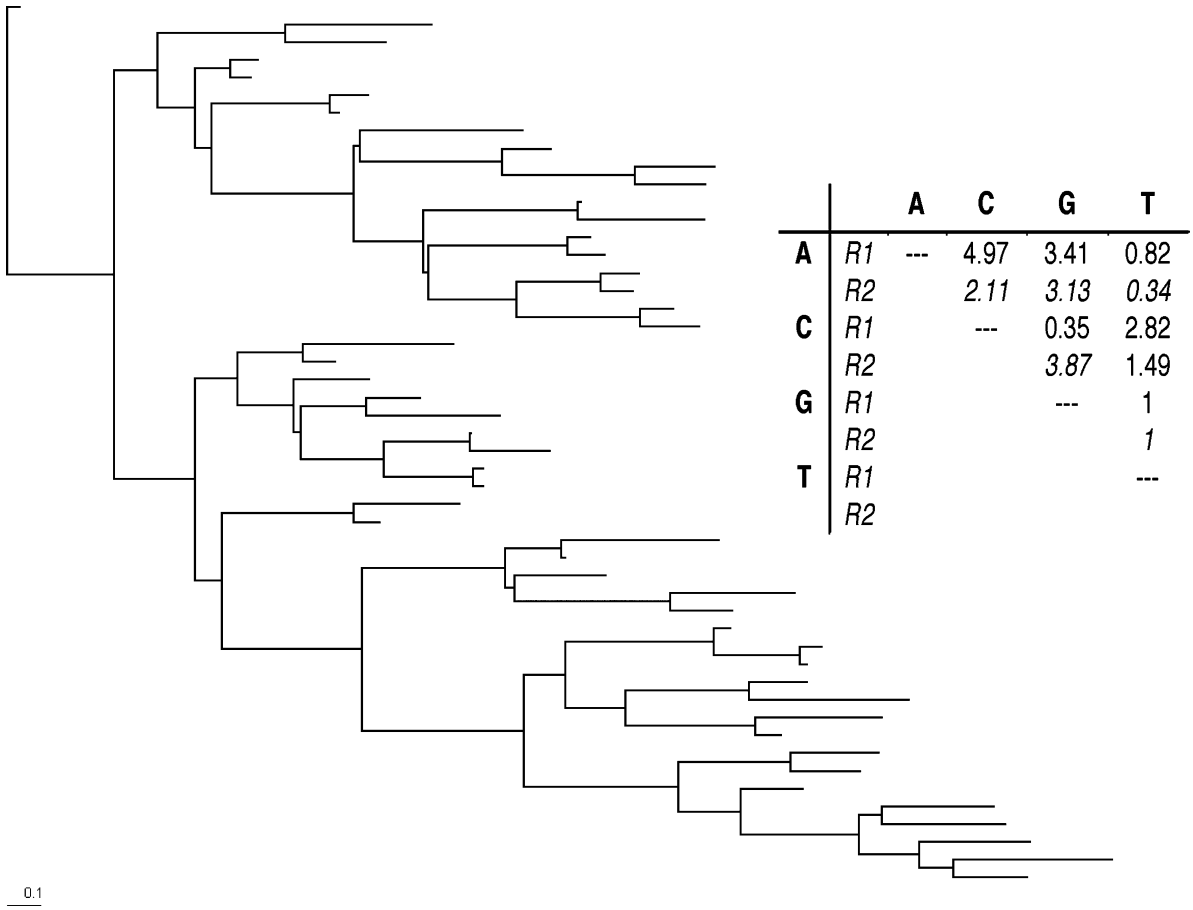
SeqGen fixes the G $\leftrightarrow$ T transition at 1.0, and so for the simulations we chose the values of the five free parameters at random from a uniform distribution on the interval of 0 to 5. This was done separately for the rate-homogeneity and gamma-rates models. We used equal base frequencies ($\pi_A = \pi_C = \pi_G = \pi_T = 0.25$). To simulate pattern-heterogeneity we reused the rate matrix from the homogeneous model and the rate matrix from the gamma-rates model, generating new data from each. One of the two was used to generate 1200 independent sites, with the remaining 800 sites coming from the second rate matrix. We simulated combined pattern and rate heterogeneity by adding the same rate heterogeneity to the two rate matrices. Figure 1 shows the tree with branch lengths, plus the instantaneous rate parameters of the two random ***R*** matrices.

RESULTS OF SIMULATIONS

We analyzed the simulated data by MCMC methods, allowing our Markov-chain to reach convergence before sampling 100 trees at widely spaced intervals (10,000 trees) to ensure independence of successive trees. We treated a chain as being at convergence when there was no average improvement in the likelihood for 200,000 iterations. We ran at least five chains for each model, and all runs converged to the same region of tree space as judged by likelihoods and posterior probabilities of trees. The means and averages we report below are based upon these samples. We used four rate categories in the gamma-rates model to analyse the data. Additional rate categories did not significantly increase the likelihood. We used two or four independent rate matrices in the pattern-heterogeneity mixture model to analyze all four simulated data sets.

Table 1 reports the average over 100 trees of the log-likelihood of each model as applied to each of the four simulated data sets. We expect that when the model used to analyze the data matches the model used to simulate the data, the fit will be adequate (entries in bold type). Simpler models are not expected to fit the data as well, and more complicated models should only lead to small improvements in the likelihood. Where models with more parameters apparently improve the fit, the Bayes factor should be small, suggesting that the improvement is only that expected given greater number of parameters, and does not signify a real difference between the two models.

The expected patterns emerge from these simulated data. The first row of the table shows that the models all perform equally well on the simulated homogeneous data (1Q), differing in their average log-likelihoods by less than 0.01%. As expected, the GTR+Γ model substantially outperforms the homogenous model when the data are generated according to a gamma-rates model.



		A	C	G	T
A	R1	---	4.97	3.41	0.82
	R2		2.11	3.13	0.34
C	R1		---	0.35	2.82
	R2			3.87	1.49
G	R1			---	1
	R2				1
T	R1				---
	R2				

FIGURE 1. Random tree of 50 tips used in the simulations. Tips are arbitrarily numbered. Branch lengths were chosen randomly from a uniform distribution on the interval 0–1. The box shows the two **R** matrices of instantaneous rates used in the simulations. These were generated at random by choosing from a uniform distribution on the interval of 0–5, with G ↔ T transitions fixed at 1.0. Matrix R1 was used to simulate the homogeneous rates model, matrix R2 was used to simulate the gamma rate-heterogeneity (using a gamma shape parameter of 1.0), and R1 and R2 were combined to generate the pattern-heterogeneity data set.

The 2Q model performs slightly worse than the GTR+Γ model on the simulated gamma data, but the 4Q model returns an approximately 23 log-unit improvement over the GTR+Γ model when both are applied to the simulated gamma data.

This result for the 4Q model confirms our expectation that the mixture model can always be made to conform to a gamma model if sufficient rate matrices are used. However, the 4Q model requires 26 parameters, compared to just 6 for the GTR+Γ model. The log-Bayes factor comparing these two models is <0, indicating that the 23 log-unit improvement of the 4Q model is expected given its greater number of parameters. The same conclusion holds for the 31 log-unit improvement of the 2Q+Γ

TABLE 1. Log-likelihoods of four different models applied to simulated gene-sequence data.

Simulation model[b]	No. of parameters	Analysis model[a] Pattern-heterogeneity				
		GTR	GTR +Γ(4)[c]	2Q	4Q	2Q+Γ
GTR	5	**−98192 (6.94)**[d]	−98193 (8.44)	−98187 (6.59)	−98191 (5.52)	−98225 (7.09)
GTR+Γ	6	−88051 (7.60)	**−82905 (6.32)**	−83782 (9.05)	−82882 (7.81)	−82874 (7.89)
2Q	12	−100864 (7.79)	−100857 (7.36)	**−100295 (7.80)**	−100294 (7.21)	−100319 (9.20)
2Q+Γ	13	−87576 (6.68)	−82556 (7.00)	−83375 (8.40)	−82506 (7.48)	**−82256 (6.56)**

[a]GTR = general time reversible model (see text). The notation for the pattern-heterogeneity model signifies that the data were analyzed using two or four independent rate matrices (see text) or two matrices plus gamma rate heterogeneity.

[b]Data were simulated from general time reversible models (GTR) with or without gamma and using pattern-heterogeneity with two rate matrices or two rate matrices plus gamma.

[c]The notation for the gamma model signifies that we used four discrete rate categories to analyze the data.

[d]Means are calculated from 100 independent MCMC trees, standard deviation in parentheses. Bold type signifies that the analysis model matches the simulation model.

TABLE 2. Input and obtained rate parameters from simulated data.

	Input and obtained rate parameters[a]						
	A ↔ C	A ↔ G	A ↔ T	C ↔ G	C ↔ T	G ↔ T	Q-weight
Rate matrix	*2.11, 4.97*	*3.13, 3.41*	*0.34, 0.82*	*3.87, 0.35*	*1.49, 2.82*	*1, 1*	*0.4, 0.6*
Q1	1.96 (0.14)	2.98 (0.2)	0.4 (0.04)	3.64 (0.26)	1.54 (0.09)	0.92 (0.09)	0.43 (0.02)
Q2	4.2 (0.27)	3.01 (0.19)	0.71 (0.07)	0.22 (0.05)	2.52 (0.15)	1 (na)[b]	0.57 (0.02)

[a]Values in italics are the transition rate parameters from the R matrix of the GTR model used to generate the simulated data. Equal base frequencies of 0.25 were assumed. Values in the main body of the table are those obtained from the 2Q pattern-heterogeneity mixture model, standard deviations in parentheses.
[b]This transition rate is fixed at 1.0.

model over the GTR+Γ model when both are applied to the simulated gamma data. By comparison the 2Q model improves upon the homogeneous and GTR+Γ models by about 560 log-units (log-Bayes factors of >500) for data generated from a 2Q model. The 2Q+Γ model applied to the simulated 2Q+Γ data substantially outperforms all the other models.

Parameter and Branch Length Estimation in Simulated Data

The 2Q mixture model accurately estimates the input transition rate parameters for the simulated 2Q data (Table 2), and returns the correct tree length (input tree length = 22.46, average obtained tree length = 23.04 ± 0.21). Figure 2 shows that the 2Q model also recovers the two different substitutional patterns on a site by site basis. The figure plots for each of the 2000 simulated sites, the difference in the log-likelihoods attributable to the two rate matrices. Positive values indicate sites for which the first rate matrix fitted the site better and vice versa. The crossover at site 1200 is where the sites began to be simulated from the second rate matrix.

We expect the 4Q model to recover transition rates from the simulated gamma data that conform to gamma expectations (Equation 2). Figure 3 plots the 24 transition rate parameters we obtained from the 4Q model against the values expected from the gamma model. The latter are obtained by multiplying the input transition rates by the four gamma scaling factors obtained from a discrete

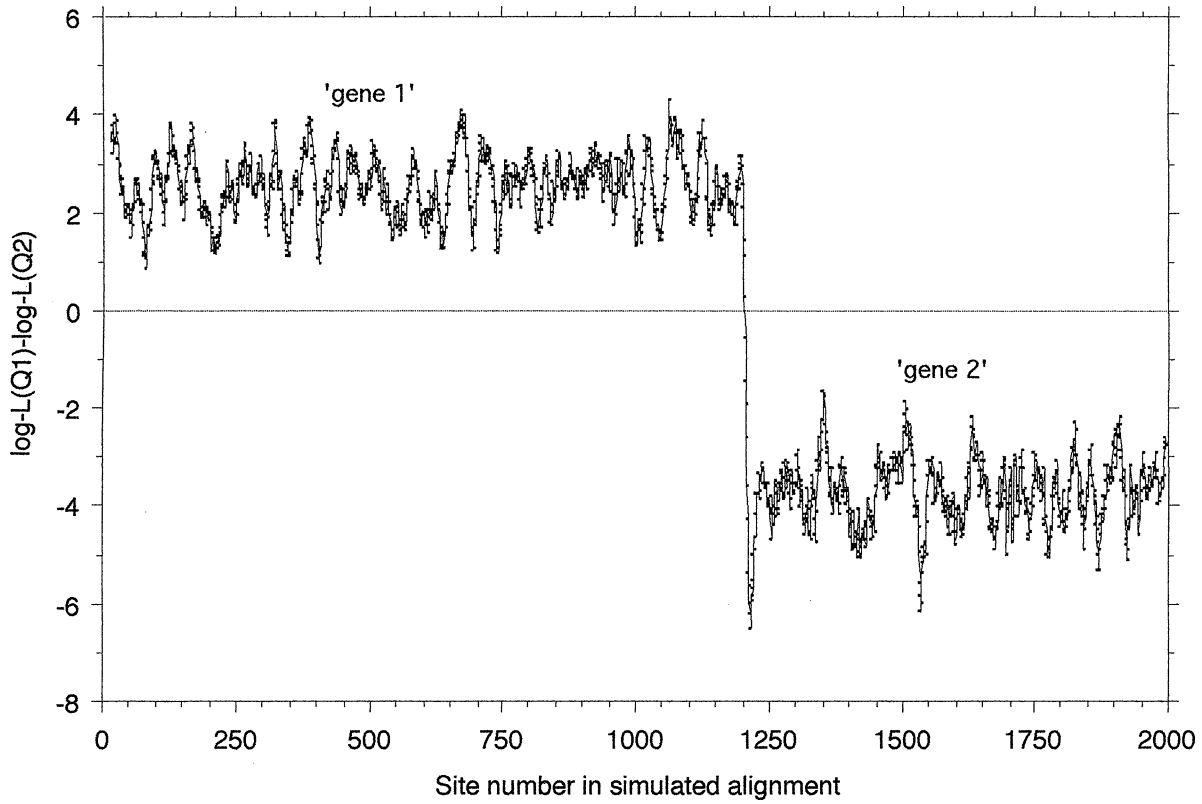


FIGURE 2. Site by site differences in the goodness of fit (log-likelihood) between the two rate matrices of the 2Q model as applied to the simulated pattern-heterogeneity data. Positive values indicate that rate matrix 1 fitted the data better than matrix 2, and vice versa. As these are logs, their difference indicates the ratio of the goodness of fit. The pattern changes at the boundary between the two simulated genes indicating that the 2Q model detects qualitatively different patterns without prior partitioning, or knowledge of the patterns in the data.

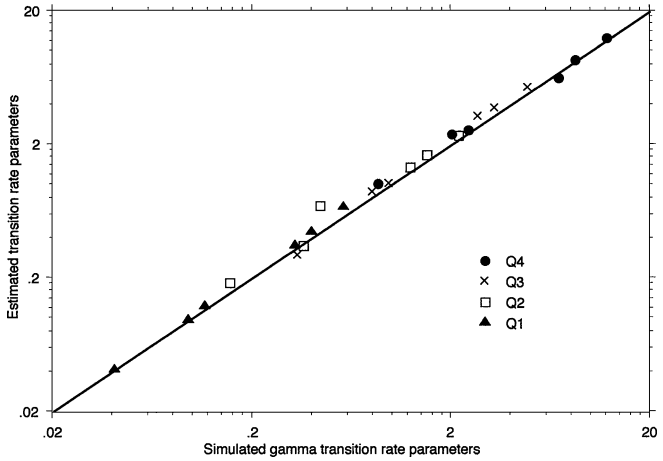


FIGURE 3. Relationship between transition rate parameters estimated by the 4Q mixture model and the input gamma-distributed transition rates used to generate the simulated rate-heterogeneity data: $r^2 = 0.996$, slope = 0.99. The legend indicates the symbols associated with each of the four rate matrices of the 4Q model. Logarithmic axes are used to improve resolution. The relationship shows that the pattern-heterogeneity model can mimic the rate-heterogeneity model.

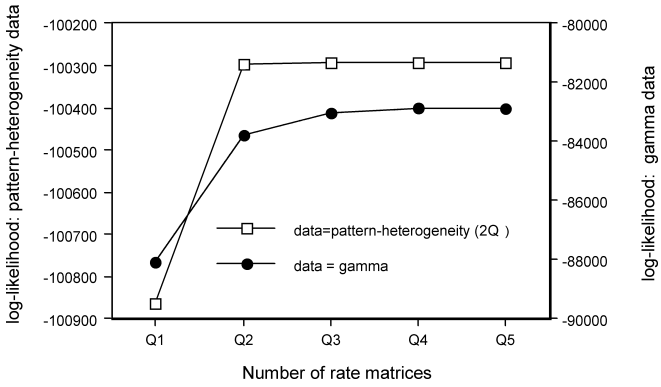


FIGURE 4. Plot of the log-likelihood of the mixture model using a range of rate matrices, applied to the pattern-heterogeneity data simulated from two rate matrices (left Y-axis), and to the gamma data simulated from a continuous gamma (right Y-axis). The expectation is that the log-likelihood should plateau at two rate matrices for the simulated pattern-heterogeneity data and at about four rate matrices for the simulated gamma data. Slight improvements with three, four, and five matrices are attributable to the greater number of parameters in these models and do not represent significant gains.

gamma distribution with shape parameter set to 1.0. The strong linear relationship between the two shows that the 4Q model is detecting the gamma signal in the data and configuring its separate rate matrices to act like the gamma model. Although this shows how gamma is a special case of the mixture model, faced with these results, one would use the gamma model to analyze these data, owing to its simplicity.

How Many Rate Matrices to Estimate?

We used a 2Q model in Figure 2 to analyze the simulated pattern-heterogeneity data, knowing that there were two patterns embedded in it. The number of rate matrices would not normally be known in advance, and so it is important to show that once the main patterns are detected, adding rate matrices does not improve the likelihood. This is analogous to determining how many rate categories to estimate under the gamma model. Figure 4 shows the likelihoods associated with the mixture model for increasing numbers of rate matrices, when applied to the simulated 2Q data and to the simulated gamma data. The likelihood reaches a plateau at two rate matrices, and for the simulated gamma data the likelihood improves very little beyond four rate matrices.

We also expect that when sufficient rate matrices have been estimated for a given data set, the parameters of additional matrices will be poorly estimated and that superfluous matrices will receive small weights (Equation 1). Figure 5 records the standard deviations of the estimated rate parameters calculated over 100 MCMC samples for the 2Q model applied to data simulated from a single rate matrix. The second matrix receives a very small weight and its parameters have very much larger standard deviations.

RESULTS IN REAL DATA

EF-1α and DDC Concatenated Alignment

Mitchell, Mitter, and Regier's (2000) concatenated alignment of the nuclear EF-1α and decarboxylase (DDC) genes for 77 noctuoid moth species totals 1949 base pairs. We analyzed these data with models using multiple rate matrices and multiple rate matrices + gamma rate-heterogeneity. The upper portion of Figure 6 plots the log-likelihoods from these two models for between one and six rate matrices (plotted data in Figure caption).

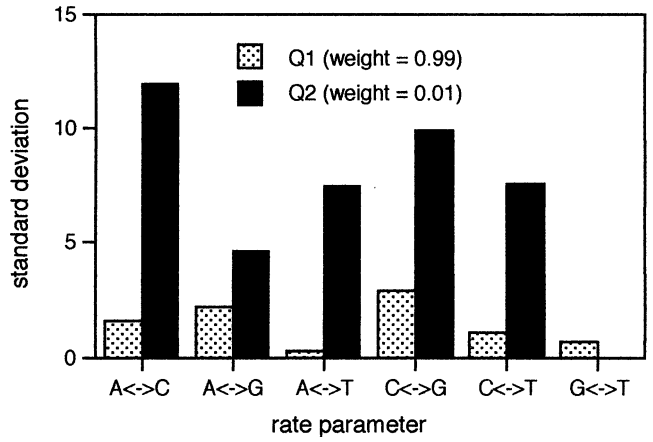


FIGURE 5. Standard deviations of the estimated rate parameters derived from the 2Q mixture model with two rate matrices applied to simulated data using one rate matrix. The model gives a weight of 0.99 to the first matrix, and the parameters are estimated within narrow limits (small standard deviations). The second rate matrix is superfluous, receives a weight of 0.01, and its rate parameters are poorly estimated. No standard deviation was calculated for the G ↔ T rate parameter for the matrix labelled Q2, it having been fixed at 1.0.

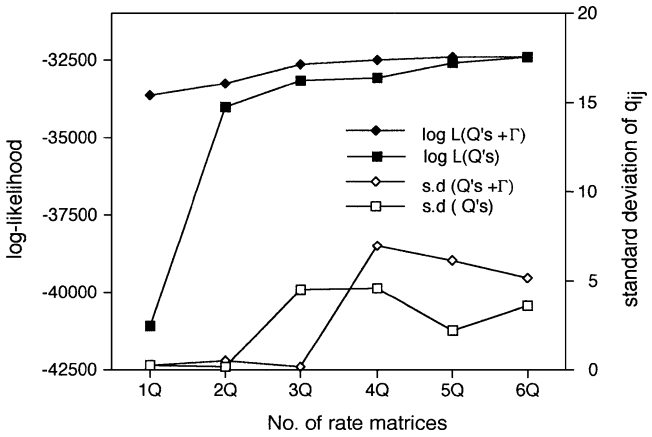


FIGURE 6. Plot of the average log-likelihoods (left-hand axis) of the mixture model with and without gamma rate heterogeneity and using a range of rate matrices, as applied to the EF-1α+DDC data (Mitchell et al., 2000). Average log-likelihoods were obtained from 100 trees sampled from the converged Markov chain. The right-hand axis records for different numbers of rate matrices the average of the standard deviations of the rate parameters. The standard deviation of each rate coefficient was measured across the same 100 MCMC samples and then averaged across all six parameters. The improvement in log-likelihood between three and four rate matrices for the combined pattern and rate heterogeneity model is 140 log-units. However, the large increase in the standard deviations of the rate coefficients at four matrices may suggest that the parameters of this model are not well estimated. Plotted data: log-likelihoods for 1–6Q model: −41084, −33998, −33129, −33060, −32556, −32395; log-likelihoods for 1–6Q +Γ model: −33614, −33219, −32607, −32466, −32410, −32389.

The comparison between the 1Q and 1Q+Γ models shows that there is a large component of rate-heterogeneity in these data. For additional rate matrices, combining rate- and pattern-heterogeneity always improves upon pattern-heterogeneity alone, except for the case of six rate matrices. At six rate matrices, the 6Q and 6Q+Γ models return virtually identical likelihoods (log L = -32395 ± 11.67 and -32389 ± 12.17, respectively). This again shows how a mixture model allowing pattern-heterogeneity can always mimic a model that includes gamma rate-heterogeneity, and here suggests that six rate matrices effectively exhaust the systematic rate- and pattern-heterogeneity variability in the data.

The question is whether a simpler model should be preferred. The 2Q+Γ improves the likelihood by 395 log-units over 1Q+Γ. The 3Q+Γ ccounts for a further 612 log-units. With a fourth rate matrix, the increase in likelihood slows to 141 log-units (log-Bayes factor > 100), a fifth matrix adds 56 log units (log-Bayes factor = 23), and the 6Q+Γ adds just 21 log units over 5Q+Γ (log-Bayes factor < 0). Superficially, these results point to the 5Q+Γ model. However, the data set has a large number of invariant sites, meaning that sites may not be independent as presumed by the Bayes factor test. There is also a large increase in the standard deviations of the estimated rate coefficients (measured in 100 MCMC samples; lower portion of Fig. 6) beginning with the 4Q+Γ model. Putting these two points together may suggest that the 3Q+Γ odel is preferred. This cut-off point corresponds to a slowing in the improvements to the overall log-likelihood of the data from the combined model. Rate matrices beyond three may begin to account for rate-heterogeneity, a hunch that is backed up by the convergence of the pattern-heterogeneity and pattern-heterogeneity + gamma models.

Adopting the 3Q+Γ model for these data gives a log-likelihood of -32607 ± 10.68, based on 20 parameters. This improves by 574 log-units on the −33181 that Mitchell et al. (2000) report for a GTR+Γ+I (invariant sites) model, having partitioned the data and fitted the model separately to each gene. Their GTR+Γ+I model also requires 20 parameters (10 rate parameters, 6 base frequencies, 2 gamma shape parameters, and 2 invariant sites parameters). Another comparison to the mixture model is to partition the data by codon position, fitting separate GTR+Γ models within each partition. This returns a likelihood of $-32,993 \pm 10.40$, and requires 27 parameters (15 rate parameters, 9 base frequencies, and 3 gamma shape parameters). The 3Q+Γ mixture model improves upon this model by 386 log-units, while using 7 fewer parameters.

Table 3 analyzes the contribution of the 3Q+Γ model's three rate matrices by studying the number of sites that each rate matrix fits best, broken down by codon position and gene in the alignment, and summed over rate categories of the gamma model. Each rate matrix specializes on a particular codon position. This is independent of variation among codon positions in rates of evolution as that is accounted for by the gamma rate heterogeneity component of the model. However, despite specializing, each matrix also provides the best fit to a large number of other codon sites. Although there are patterns that statistically distinguish the codon positions, they are not unique to those positions. This is why partitioning the data by gene or by codon position performs worse that fitting the mixture model: partitioning misses the within-partition variability.

Analysis of Patterns in 12S Ribosomal RNA

The stem and loop secondary structure of ribosomal RNA is predicted to be associated with different patterns of evolution. Stems or helices of ribosomal RNA

TABLE 3. Numbers of sites for which a given rate matrix fits the data best, broken down by codon position: EF-1α+DDC data.[a]

	Gene 1 Codon position			Gene 2 Codon position		
Rate matrix	1	2	3	1	2	3
Q1	282	197	172	117	96	61
Q2	126	214	111	77	116	33
Q3	5	2	131	42	24	143

[a]The data consist of a concatenated alignment of two genes and 1949 sites: EF-1α has 1240 sites and the decarboxylase gene has 709 sites. The data were analyzed using three independent rate matrices plus gamma rate variability across all sites (see text, Equation 2). The number in each cell records the number of sites at a given codon position for which the corresponding rate matrix provided the best fit.

TABLE 4. Log-likelihoods of models fitted to 12S data, $n = 54$ mammal species.

Model[a]	Log-likelihood mean (SD)[b]	
GTR	−27097 (9.80)	
GTR+Γ (4 categories)	−23233 (9.97)	
GTR+Γ (partitioned)[c]	−22790 (6.33)	
	2 Q's	4 Q's
Pattern-heterogeneity[d]	−23772 (9.44)	−22981 (8.47)
Pattern-heterogeneity + Γ[e]	−22914 (9.23)	−22730 (7.37)

[a]GTR = general time reversible model (see text).
[b]Means calculated from 100 independent MCMC trees, standard deviation in parentheses.
[c]Data partitioned and separate GTR+Γ models fitted to stem versus loop sites.
[d]Data were analyzed using either two (2Q) or four (4Q) rate matrices, with and without gamma rate heterogeneity.
[e]Base frequencies (π_i) allowed to vary among Q matrices owing to large differences between stems and loops.

often adopt canonical Watson-Crick base pairing whereas loops are unpaired. Nucleotide substitutions at a given site in a stem are therefore expected to be compensated by a change at the paired site. Higgs (1998) suggests that this will lead to high transition/transversion ratios in stems. No predictions are made for loops.

We fitted the mixture model with and without rate-heterogeneity to the 12S ribosomal DNA data from the 54 mammal species Jow et al. (2003) report. We also fitted a model in which the data were partitioned into stem versus loop sites. The combined 4Q+Γ model improves upon the 2Q+Γ model by 184 log-units (Table 4; log-Bayes factor of 119.5). This example, like that above for the two protein coding genes, shows that there is substantial pattern-heterogeneity in these data even after accounting for the rate-heterogeneity.

Table 5 reports the rate parameters obtained from fitting four rate matrices to the 12S data. Rate matrix Q4 has the highest ratio of transitions to transversions and recovers the predicted pattern for stem substitutions (high Tr/Tv and slow-fast-slow-slow-fast-slow ordering of transition rates as listed in Table 5). Matrix Q3 also shows this pattern but seems to identify sites in which transversions occur at higher rates than in Q4. Matrices Q1 and Q2 do not show the predicted stem pattern.

Table 6 analyzes the fit of these four rate matrices to stems and loops using the predicted secondary structure

TABLE 5. Transition rate parameters for mammalian 12S data using four rate matrices.[a]

Rate matrix[b]	Transition rate parameters						
	A ↔ C	**A ↔ G**	A ↔ T	C ↔ G	**C ↔ T**	G ↔ T	Tr/TV
Q1	15.34	**2.91**	13.00	1.50	**95.32**	1	5.02
Q2	27.64	**18.42**	19.71	6.62	**27.76**	4.93	1.53
Q3	4.94	**25.24**	3.79	1.40	**30.08**	3.49	7.66
Q4	0.18	**18.88**	0.48	0.82	**14.33**	0.30	41.15

[a]Transitions **A ↔ G and C ↔ T** in bold, transversions in regular type. Values are from the **R**-matrix, that is, the matrix of instantaneous rates, not scaled by the base frequencies. The G ↔ T rate of 1.0 in Q1 is fixed in advance. Rate matrix Q3 shows the high transition-transversion ratio and "slow, fast, slow, slow, fast, slow" pattern of transition rates predicted to hold for compensatory substitution in stems of ribosomal molecules.
[b]Independent rate matrices used in the pattern-heterogeneity model.

TABLE 6. Numbers of stem and loop sites fitted best by respective rate matrices: mammalian 12S data.

Secondary structure	Q1	Q2	Q3	Q4
Stem	76	21	71	296
Loop	133	112	83	249

of 12S (Springer and Douzery, 1996; P. Higgs, personal communication). Matrices Q1 and Q2 seem to characterize loop evolution. Matrices Q3 and Q4 fit the majority of stem sites best, but also fit a large number of loop sites. In fact, the matrix Q4 emerges as a generalist matrix, roughly evenly divided between fitting stems and loops.

DISCUSSION

The pattern-heterogeneity mixture model is a general tool that can be applied to any kind of aligned data set, including proteins or morphological traits. Our simulation studies show that it retrieves the model of sequence evolution used to produce the data, and outperforms either homogeneous or rate-heterogeneity models when rate- or pattern-heterogeneity are present. The model is useful for analyzing patterns of evolution within a single gene, and can be applied to data sets derived from more than one gene. An attractive feature of the model is that the investigator can observe how different rate matrices may characterise different regions of the data, without partitioning it. We have not given examples of applying the mixture model to morphological traits, but have implemented a model to do so (Pagel, 1994; Lewis, 2001) and find that morphological traits can also exhibit pattern-heterogeneity.

Our analysis of the combined EF-1α and DDC data (Mitchell et al., 2001) showed how a mixture model can improve upon partitioning. Partitioning the data by gene or by codon position returned poorer likelihoods than allowing the mixture model to settle on three rate matrices. These two protein coding genes exhibited different patterns of evolution, and they showed the characteristic differences at first, second, and third codon positions. However, the evolutionary variability within genes and within codon position is large and this is missed by the partitions. By comparison the mixture model finds three rate matrices that exhibit some specialisation on codon position but seem also to capture other patterns not obvious from mere inspection of the data.

In our application of the model to mammalian ribosomal 12S data, we were able to identify specific rate matrices for stems and loops. These show that the pattern of evolution for a majority of the stem sites does not conform to the predicted 'stem' rate matrix (see also Hickson et al., 1996; Simon et al., 1994). Here again, partitioning would miss this variability. The patterns in the mixture model are emergent, not being specified or constrained in any way a priori and not being dependent upon knowledge of the secondary structure. They could be of importance in predicting secondary structure, and by implication, in understanding the behaviour of the secondary structure of the molecule.

Methods developed especially for ribosomal data apply specific models of paired-sites sequence evolution to stems, ignoring loop data (Schöniger and von Haeseler, 1994; Savill et al., 2001). This approach assumes that the secondary structure is accurate as the coordinates of the stems and loops are supplied in advance by the investigator, and that all sights of a particular structure follow a single specific model. Its strengths are that it allows the investigator to test specific hypotheses based on theoretically justified models of evolution. The mixture model alternative is to fit separate rate matrices to the data, without any prior expectation as to their form or what sites they may fit, allowing the data to reveal its patterns of evolution.

Applied to multiple-gene concatenated data sets, a mixture model does not require that any more parameters be estimated than with partitioned data. The difference is that with the mixture model all of the data are used to estimate each parameter, and rate matrices are free to fit more than one class of site. Even where the primary variation amongst sites is in rates of evolution, it may not conform to gamma rate heterogeneity. In such cases a mixture model of pattern-heterogeneity can characterize the rate variability properly and lead to substantial improvement over the gamma model. However, when rates of evolution do conform to a gamma distribution, there will be little improvement from the mixture model and the gamma model should be used in preference to it, being specified by fewer parameters.

Mixture models can be difficult to fit to data, owing to the large number of parameters. This is largely a technical issue and not a function of the model per se. But it means that the model is best applied to larger data sets, and we suggest as a general rule of thumb in any phylogenetic inference, 10 data points per parameter. This can be estimated as *sites* × *species* or more generally as *operational taxonomic units* × *characters*, but bearing in mind that neither sites nor species are likely to be independent. A better estimate might be *number of different characters* × *number of OTUs*.

Even allowing for these very rough and probably liberal estimates, a given data set may not support two or more models of evolution (these cautionary remarks apply to partitioning or use of the gamma rate-heterogeneity model as well as to the pattern-heterogeneity model). Indications of this are a very small estimated weight (see Equation 1) for a rate matrix or large variability in the estimates of the rate parameters. In some cases, for example when a given class of data has been analyzed many times such as with HIV or flu data, one may have informed prior expectations about the number of rate matrices to fit.

Our use of a Bayes factor approach for choosing among models is, like the use of the likelihood ratio statistic in maximum likelihood (e.g., Posada and Crandall, 2001), dependent upon a number of assumptions, particularly that characters in the alignment are independent. In our own work we look for 70 to 80 log-units as a minimum contribution from an additional GTR matrix of six parameters plus one weight parameter (corresponding here to a log-Bayes factor using wide and uninformative priors of about 40).

Probable lack of independence among sites provided an additional reason for ignoring the relatively large log-Bayes factors we obtained for some of the model comparisons using the EF-1α+DDC data. For ribosomal data, one should additionally bear in mind that if stem sites are truly compensatory in their evolution, then the number of independent stem sites is roughly half of the total number of stem sites. It may be useful to apply Bayes factors in conjunction with a more subjective 'scree' test. This test plots the overall likelihood against the number of rate matrices fitted to the data, looking for an obvious turning point where the rate of increase in likelihood with additional matrices slows greatly. Analyzing the variation in fitted rate parameters may also be useful.

The results we have reported for the pattern-heterogeneity mixture model send the encouraging message that phylogenetically structured data harbor complex signals of the history of evolution, and that it is possible to design general models to detect those signals. To the extent that these signals are not lost or overwritten by more recent evolutionary events, investigators can use statistical approaches validly to infer the nature and modes of past evolutionary events and processes (Pagel, 1999), complementing experimental and palaeontological methods. We have implemented the mixture model in a computer program available from http://www.ams.reading.ac.uk/zoology/pagel. It uses an MCMC framework for inferring trees and estimating parameters of the models of evolution for sequence or morphological data.

ACKNOWLEDGMENTS

Preliminary drafts of this work were presented at the workshop on Mathematical and Computational Aspects of the Tree of Life at the Center for Discrete Mathematics and Computer Sciences (DIMACS) at Rutgers University in March, 2003, and at the workshop on the Mathematics of Evolution and Phylogeny, Institute Henri Poincaré, Paris, June, 2003. This work is supported by grants 45/G14980 and 45/G19848 to MP from the Biotechnology and Biological Sciences Research Council (UK).

REFERENCES

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

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 1995. Pages 420–438 *in* Bayesian data analysis. Chapman and Hall, London.

Geyer, C. J. 1992. Practical Markov chain Monte Carlo. Stat. Sci. 7:473–511.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter. 1996. Introducing Markov chain Monte Carlo. Pages 1–19 *in* Markov chain Monte Carlo in practice (W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, eds). Chapman and Hall, London.

Hickson, R. E., C. Simon, A. Cooper, G. S. Spicer, J. Sullivan, and D. Penny. 1996. Conserved sequence motifs, alignment, and secondary structure for the third domain of animal 12S rRNA. Mol. Biol. Evol. 13:150–169.

Higgs, P. G. 1998. Compensatory neutral mutations and the evolution of RNA. Genetica 102:91–101.

Hillis, D. M., and M. T. Dixon. 1991. Ribosomal DNA: Molecular evolution and phylogenetic inference. Q. Rev. Biol. 66:411–453.

Huelsenbeck, J. P., and Nielsen, R. 1999. Variation in the pattern of nucleotide substitution across sites. J. Mol. Evol. 48:86–93.

Huelsenbeck, J. P., F. Ronquist, R. Nielsen, and J.P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310–2314.

Jow, H., C. Hudelot, M. Rattray, and P. G. Higgs. 2002. Bayesian phylogenetics using an RNA substitution model applied to early mammalian evolution. Mol. Biol. Evol. 19:1591–1601.

Koshi, J. M., and R. A. Goldstein. 1998. Models of natural mutations including site heterogeneity. Proteins: Struct. Funct. Genet. 32:289–295.

Krajewski, C., M. G. Fain, L. Buckley, and D. G. King. 1999. Dynamically heterogeneous partitions and phylogenetic inference: An evaluation of analytical strategies with cytochrome b and ND6 gene sequences in cranes. Mol. Phylo. Evol. 13:302–313.

Larget, B., and D. L. Simon. 1999. Markov chain monte carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol. Biol. Evol. 16:750–759.

Lartillot, N., and H. Philippe. 2004. A Bayesian mixture model for across site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. in press.

Lewis, P. O. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. Syst. Biol. 50:913–925.

Lutzoni, F., M. Pagel, and V. Reeb. 2001. Major fungal lineages derived from lichen-symbiotic ancestors. Nature 411:937–940.

Mitchell, A., C., Mitter, and J.C. Regier. 2000. More taxa or more characters revisted: Combining data from nuclear protein-coding genes for phylogenetic analyzes of Noctuoidea (Insecta: Lepidoptera). Syst. Biol. 49:202–224.

Murphy, W. J., E. Eizirik, S. J. O'Brien, O. Madsen, M. Scally, C. J. Douady, E. Teeling, O.A. Ryder, M. J. Stanhope, W. W. de, Jong, and M. S. Springer. 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science 294:2348–2351.

Pagel, M. 1994. Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. Proc. R. Soc. B 255:37–45.

Pagel, M. 1999. Inferring the historical patterns of biological evolution. Nature 401:877–884.

Pagel, M., and F. Lutzoni. 2002. Accounting for phylogenetic uncertainty in comparative studies of evolution and adaptation. Pages 148–161 *in* Biological evolution and statistical physics (M. Lässig and A. Valleriani, eds.). Springer-Verlag, Berlin.

Posada, D., and K. A. Crandall. 2001. Selecting the best fit model of nucleotide substitution. Syst. Biol. 50:580–601.

Rambaut, A., and N. C. Grassly. 1997. Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:235–238.

Raftery, A. E. 1996. Hypothesis testing and model selection. Pages 163–188 *in* Markov chain Monte Carlo in practice (W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, eds.). Chapman and Hall, London.

Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804.

Savill, N. J., D. C. Hoyle, and P. G. Higgs. 2001. RNA sequence evolution with secondary structure constraints: Comparison of substitution rate models using maximum likelihood methods. Genetics 157:399–411.

Schöniger, M., and von A. Haeseler. 1994. A stochastic model for the evolution of autocorrelated DNA sequences. Mol. Phylo. Evol. 3:240–247.

Simon, C., F. Frati, A. Beckenbach, B. Crespi, H. Liu, and P. Flook. 1994. Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. Ann. Entomol. Soc. Am. 87:651–701.

Springer, M. S., and E. Douzery. 1996. Secondary structure and patterns of evolution among mammalian mitochondrial 12S rRNA molecules. J. Mol. Evol. 43:357–373.

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. Pages 407–514 *in* Molecular systematics, 2nd edition (D. M. Hillis, C. Moritz, and B. K. Mable, eds.). Sinauer, Sunderland, Massachusetts.

Wilson, I., and D. Balding. 1998. Genealogical inference from microsatellite data. Genetics 150:499–510.

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

Yang, Z., R., Nielsen, N., Goldman, and A.-M. K. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431–449.

First submitted 30 June 2003; reviews returned 23 October 2003; final acceptance 29 January 2004
Associate Editor: Keith Crandall