BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 19 2008, pages 2177–2183
doi:10.1093/bioinformatics/btn395

Phylogenetics

Site-specific evolutionary rates in proteins are better modeled as non-independent and strictly relative

Andrew D. Fernandes[1,2,3,4,5,*] and William R. Atchley[3,4,5]

[1]Department of Biochemistry, The University of Western Ontario, London, Ontario, N6A 5C1, [2]Department of Applied Mathematics, The University of Western Ontario, London, Ontario, N6A 5B7, Canada, [3]Graduate Program in Biomathematics, North Carolina State University, Raleigh, NC 27695-8203, [4]Center for Computational Biology, North Carolina State University, Raleigh, NC 27695-7614 and [5]Department of Genetics, North Carolina State University, Raleigh, NC 27695-7614, US

Received on May 2, 2008; revised on July 23, 2008; accepted on July 25, 2008

Advance Access publication July 28, 2008

Associate Editor: Martin Bishop

ABSTRACT

Motivation: In a nucleotide or amino acid sequence, not all sites evolve at the same rate, due to differing selective constraints at each site. Currently in computational molecular evolution, models incorporating rate heterogeneity always share two assumptions. First, the rate of evolution at each site is assumed to be independent of every other site. Second, the values of these rates are assumed to be drawn from a known prior distribution. Although often assumed to be small, the actual effect of these assumptions has not been previously quantified in the literature.

Results: Herein we describe an algorithm to simultaneously infer the set of $n-1$ *relative* rates that parameterize the likelihood of an n-site alignment. Unlike previous work (a) these relative rates are completely identifiable and distinct from the branch-length parameters, and (b) a far more general class of rate priors can be used, and their effects quantified. Although described in a Bayesian framework, we discuss a future maximum likelihood extension.

Conclusions: Using both synthetic data and alignments from the Myc, Max and p53 protein families, we find that inferring *relative* rather than *absolute* rates has several advantages. First, both empirical likelihoods and Bayes factors show strong preference for the relative-rate model, with a mean $\Delta \ln P = -0.458$ per alignment site. Second, the computed likelihoods and Bayes factors were essentially independent of the relative-rate prior, indicating that good estimates of the posterior rate distribution are not required a priori. Third, a novel finding is that rates can be accurately inferred even when up to $\approx$4 substitutions per site have occurred. Thus biologically relevant putative hypervariable sites can be identified as easily as conserved sites. Lastly, our model treats rates and tree branch-lengths as completely identifiable, allowing for the first time *coherent* simultaneous inference of branch-lengths and site-specific evolutionary rates.

Availability: Source code for the utility described is available under a BSD-style license at http://www.fernandes.org/txp/article/9/site-specific-relative-evolutionary-rates.

Contact: andrew@fernandes.org

Supplementary information: Supplementary data is available at *Bioinformatics* online.

*To whom correspondence should be addressed.

1 BACKGROUND

In a nucleotide or amino acid sequence, the rate of evolution at a given site is expected to vary according to the specific selective constraints at that site. Thus we expect a priori that not all sites evolve at the same rate (Corbin and Uzzell, 1970). Sites that are under strong selective constraints should be relatively highly conserved, while sites under lesser selective pressure should be more variable. In essence, the observed evolutionary rate corresponds to the level of purifying selection at that site (Kimura, 1983). We know that, in a phylogenetic analysis, not accounting for this rate heterogeneity can yield misleading results (Felsenstein, 2001; Yang, 1994, 1996; Yang and Kumar, 1996). Therefore, correctly modeling rate heterogeneity is important both for correct phylogenetic reconstruction and the discrimination of conserved from non-conserved sites.

Traditionally, given data D consisting of a fixed n-site alignment and tree topology, the likelihood of observing D given rates $r = [r_1, r_2, \ldots, r_n]$ and branch-lengths $t = [t_1, t_2, \ldots, t_m]$ is

$$P(D|r,t) = \prod_{i=1}^{n} P_i(r_i t), \tag{1}$$

where P_i denotes the likelihood of site i. In this model the likelihood of each site is independent of every other site. Furthermore, the set of rates r and branch-lengths t are not completely identifiable for any dataset because $P_i(r_i t) = P_i\left(r_i s^{-1} \cdot st\right)$ for any $s > 0$. In other words, halving the rates and doubling the branch-lengths yields the same likelihood.

One of the first attempts to use one rate per site to estimate the overall likelihood was made by Swofford *et al.* (1996), using maximum likelihood, in the DNARATES program. However, Felsenstein (2001, 2004) subsequently cautioned that the 'one rate-parameter per site' model may lead to an ill-conditioned maximum likelihood model since the number of model parameters increases linearly with the number of alignment sites.

To regularize the likelihood calculation (1), Uzzell and Corbin (1971), followed by Nei *et al.* (1976), assumed that rather than being fixed, each rate was drawn from a known prior distribution. Each rate was further assumed independent of every other rate. The likelihood of each site could then be integrated *independently* over

all possible rates. More formally, they calculated

$$P(D|r,t) = \prod_{i=1}^{n} \left[\int_{\overline{\mathbb{R}^+}} f(r_i) P_i(r_i t) \, dr_i \right], \quad (2)$$

where $f(r_i)$ denotes the density of the rate prior and $\overline{\mathbb{R}^+}$ denotes the non-negative reals. The unit mean gamma distribution was historically used for f because it often yields analytically tractable models. For calculations that are less amenable to analytic results, the discrete gamma approximation, first popularized by Yang (1994), has become the *de facto* standard rate prior in molecular evolution.

Unfortunately, the assumptions inherent in (2) result in two undesirable, yet unavoidable, consequences. First, enforcing a unit mean constraint on the prior f does not constrain the posterior rates in any useful manner, as can be seen in sample calculations from recent versions of PHYML (Guindon and Gascuel, 2003) or MRBAYES (Ronquist and Huelsenbeck, 2003). Thus rates and branch lengths remain mathematically unidentifiable in this model. The RATE4SITE program by Pupko *et al.* (2002) takes the re-normalization approach suggested by Meyer and von Haeseler (2003) whereby rates and branch lengths are estimated by alternately inferring rates given the lengths, then the lengths given the rates. At each step, the rates are re-normalized to have a unit mean. The consequences of inferring rates (and phylogeny) without rigorously dealing with rate/time non-identifiability has not been quantified or formally investigated. A more detailed discussion of this issue can be found in the Supplementary Material.

The second undesirable consequence inherent in (2) is that the actual distribution f must be either specified or estimated. Most frequently, a unit-mean gamma distribution is assumed, and the shape parameter α of that distribution is estimated. Several attempts at addressing the shortcomings of the unit-mean gamma rate prior have been undertaken, most notably with Gu *et al.* (1995) who augmented the gamma distribution with an estimated proportion of invariant sites (where $r_i = 0$). Mayrose *et al.* (2005a) advocated using a mixture of gamma distributions, while Pond and Frost (2005) used more general parameterized distributions.

Again, the consequences of inferring rates (and phylogeny) under the influence of these priors is not known. For instance, we may compare a model using one gamma prior with another using a two-gamma mixture. If the two-gamma mixture does not yield a significantly better model, we may erroneously conclude that the single-gamma model is a 'good' approximation of the 'true' set of rates. In fact, this observation only supports the conclusion that under the class of n-gamma mixture priors, $n = 1$ is sufficient. For instance, since all gamma distributions have exponentially decreasing tails, this class of priors does not include models with heavy tails. To properly assess the effect of the prior, the class of rate priors should ideally be as large (in some sense) as possible.

In order to quantify the impact of (2) on rate inference, we inferred site-specific rates for both synthetic data and alignments from the Myc, Max and p53 protein families. Specifically, we infer rates $r = [r_1, r_2, \ldots, r_n]$ under the constraint that

$$\frac{1}{n} \sum_{i=1}^{n} r_i = 1. \quad (3)$$

Computationally, (3) is much more stringent than the constraint that the distribution f in (2) have unit mean. The constraint lets us model the n rates via $n-1$ *relative*-rate parameters. Relative rates are advantageous compared to absolute rates because relative rates are completely identifiable from branch lengths in likelihood calculations. This advantage comes at a price, however, in that it becomes non-trivial to integrate likelihoods over the space of all rates, subject to constraint (3).

As written, constraint (3) implies that rates are modeled as fixed-effects, and not the random-effects model more commonly assumed by traditional maximum likelihood models. If a random-effects is preferred, (3) could be re-written to imply that $\sum_i r_i \sim M(\mu)$ for some distribution M with mean rate μ. In doing so, however, we lose identifiability between rates and branch-lengths, and drastically reduce stability and convergence rate of our algorithm (data not shown).

2 RESULTS AND DISCUSSION

A Markov Chain Monte Carlo (MCMC) approach was used to integrate (1) under constraint (3) over all possible relative rates. Both simulated and real data were used to compare our relative-rate model with the best unit-mean gamma, independent-rate model. A Bayesian framework was adopted for three reasons. First, previous works suggested that empirical Bayesian methods were significantly better than likelihood methods when inferring site-specific rates (Mayrose *et al.*, 2004). Second, unlike the independent-rate assumption, constraint (3) precludes the use of simplifying numeric approaches such as Gaussian quadrature (Fernandes and Atchley, 2006) to integrate over all possible rates. Lastly, since the relative-rate model is not nested in the absolute-rate model, comparing their model fits via likelihood is not trivial. Instead, Bayes factors (Kass and Raftery, 1995), which are implicitly correct for differences in parameter dimension, are used for comparison. Throughout, we have assumed without loss of generality that branch lengths are fixed while inferring rates. Since rates and branch lengths are completely independent in our model, it is implicit that lengths could be simultaneously inferred in parallel with rates.

2.1 Rate priors

Our method is based on Bayesian techniques and thus requires specification of a relative-rate prior distribution; we assume implicitly that parameters must have well-defined posterior sampling distributions. As we will discuss, this prior is markedly different than those used for absolute-rate models. Furthermore, a rate prior is *also required* for maximum likelihood inference. To see why, recall that as long as a site is not completely conserved, $P_i(r_i t)$ approaches a positive, non-zero constant as $r_i t \to \infty$. Thus if $f(r_i)$ in (2) was constant, the integral of their product would be infinite. In fact, the likelihood $P_i(r_i t)$ is, in general, not a density with respect to r_i. Therefore, even in a maximum likelihood setting a rate prior is required to regularize the likelihood function. Non-integrable likelihood functions can sometimes be regularized with straightforward methods, such as in the case of Gaussian mixture models (Wasserman, 2000). Unfortunately, under the independent-rates assumption, such regularizations are not possible. Furthermore, it is difficult to quantify the precise effect that a family of priors will have on the final inference. For more discussion of this topic, see the Supplementary Material.

Often the required regularization constraint is 'hidden' within a method. For example, a well known early study by Kelly and Rice (1996) describes a purportedly 'priorless' rate inference procedure. In reality, their posterior rate distribution is estimated by using the moment generating function, which in itself is estimated through the eigen values of the infinitesimal rate (mutation) matrix. However, since the rate matrix itself is constrained to have have a unit-mean rate, the moments of their posterior rate distribution are automatically *implicitly* constrained, analogously to constraint (3).

Although seemingly a subtle change, the constraint (3) changes the situation significantly. Rather than integrating over the infinite domain $r \in \overline{\mathbb{R}}^{n+}$ (the n-dimensional orthant of non-negative reals), we now integrate over the finite domain $r \in \mathbb{S}^n$, the $(n-1)$-dimensional unit simplex. Examples of familiar, low-dimensional simplexes are shown in Supplementary Figure A1. The non-informative, and in this case maximum entropy, prior $f(r_i)=1$ becomes perfectly admissible. Such a simple prior may not be the optimal choice, however; there is tremendous literature describing the selecting priors based on systematic and formal rules (Berger, 2006; Kass and Wasserman, 1996). Denoting $\theta_i = r_i/n$, as the scaled relative rate, then $P(\theta) = (\theta_1\theta_2\cdots\theta_n)^{-\delta}$, $\delta \in [0,1)$ are the most common priors over the domain $\mathbb{S}^n$. $\delta = 1/2$ yields Jeffreys' prior (Jeffreys, 1946), while $\delta \to 1$ yields Jaynes' invariant Haar-measure prior (Jaynes, 1968; Syversveen, 1998). Unfortunately, both of these priors are based on examination of the multinomial likelihood function and are not appropriate for inferring rates. For instance, they imply that $r_i \to n$ is just as probable as $r_i \to 0$, even though it is biologically assumed that very high mutation rates (hundreds of times the mean rate) are quite unlikely. In fact, we found that all formulaic recipes for the construction of objective priors (Berger, 2006; Bernardo and Ramon, 1998; Bernardo and Smith, 1994; Kass and Wasserman, 1996) failed when applied to phylogenetic likelihoods since these likelihoods (a) are not densities with respect to r, assuming independence or (b) have variance increasing linearly with n, assuming relative rates.

Therefore, we chose to investigate inferential differences resulting from the use of two different priors based on intuitively reasonable assumptions. First, the uniform prior $P(r_i) \propto 1$ was selected as an appropriate comparison for a 'prior-less' maximum likelihood-type situation. Second, the unit-exponential $P(r_i) \propto \exp(-r_i)$ was selected to represent the idea that very high substitution rates are anticipated to be unlikely. Note that the relative-rate unit-exponential prior is *not* conceptually or computationally identical to assuming $f(r_i) = \exp(-r_i)$ in (2) due to the action of constraint (3).

2.2 Simulation study

To assess the behavior of our method, we inferred the rates of a synthetic dataset designed to mimic an experimentally ideal situation. Our synthetic dataset was comprised of 100 sequences of 2000 sites with no gaps. All descendants were taken to be $t=1$ time-units away from the ancestor, and the ancestral sequences were drawn from the WAG (Whelan and Goldman, 2001) equilibrium density. Rates were equally log-spaced from 10^{-3} to just under 10, with a mean of exactly 1. The prior was unit-exponential per site. A box-plot of the posterior rate distributions is shown in Figure 1. The solid sigmoidal curve denotes the site rate mean, smoothed across adjacent sites. The dotted line denotes the original rate of the simulation. Although not shown, virtually identical results were attained under the uniform prior, with no discernible qualitative differences between plots.

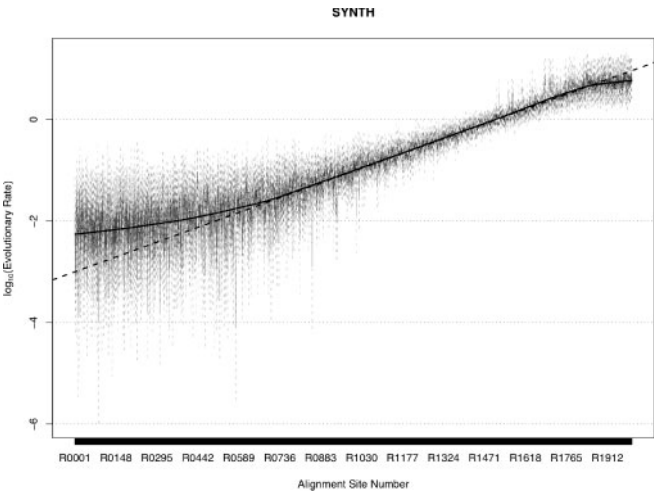

Fig. 1. Synthetic data were comprised of 100 sequences of 2000 sites with no gaps. All descendants were taken to be $t=1$ time-units away from the ancestor, and the ancestral sequences were drawn from the WAG equilibrium density. Rates were equally log-spaced from 10^{-3} to just under 10, with a mean of exactly 1. The prior was unit-exponential per rate. The solid sigmoidal curve denotes the site rate mean, smoothed across adjacent sites. The dotted line denotes the original rate of the simulation. Although not shown, virtually identical results were attained under the uniform prior.

When rates are low, few substitutions are observed, leading to two effects on inference. First, given only 100 sequences, there is no observed difference between, say, a rate of 10^{-3} and $10^{-2.3}$. At each of these rates, it is unlikely that even one substitution has occurred. Therefore, given a constraint that the mean rate equals one, highly conserved sites will have their rates biased upwards. Figure 1 shows that significant departures from mean estimated rate occur when $r_i \lessapprox 10^{-1.6} \approx 0.025$. Second, the variance of the estimated rate becomes large as the rate decreases, again as shown in the figure box-plots. This increased variance can be understood by using the analogy of estimating the rate parameter of a Poisson process when the observed event is rare. In the Poisson case, the expected Fisher information is inversely proportional to the number of events observed, which by assumption is small. Hence, the variance of the estimated rate of a conserved position will be large.

For rates between $\approx 10^{-1.6}$ and $\approx 10^{0.60} \approx 4.0$ the mean inferred rate is almost completely coincident with the actual rate. We found the magnitude of the upper bound rate (4.0) surprising, since it implied that evolutionary rates could be accurately inferred even when, on average, four substitution events occurred between every observed sequence in the test data. Prior experience with other biological datasets led us to expect that such a high substitution rate would be indistinguishable from complete randomization $(r_i \to \infty)$. Figure 1 shows that for the correct dataset there is considerable discernible difference between high substitution rates and randomization. We hypothesize that most substitution events given by amino acid evolution models substitute amino acids primarily within the same 'similarity' class; aliphatic, aromatic, charged and so on. Since estimating rates considers substitution both *within* and *between* amino acid similarity classes, with enough

Table 1. Comparison of model fit likelihoods and posterior probabilities for the independent- and relative-rate models

	Dataset	MAX		MYC		p53		p53R		Synth	
	sites	380		79		1137		718		2000	
	sequences	23		45		64		15		100	
	prior	exp(−*r*)	uniform	exp(−*r*)	uniform	exp(−*r*)	uniform	exp(−*r*)	uniform	exp(−*r*)	uniform
Model comparison	Max ln L	-4538.804	−4538.804	−2076.357	−2076.357	−33756.110	−33756.110	−6815.064	−6815.064	−227602.474	−227602.474
	ln P(M)	−4396.984	−4396.533	−2029.543	−2029.312	−33133.550	−33133.818	−6588.002	−6587.548	−224230.494	−224241.380
	Max ln P	−4340.639	−4332.986	−2000.333	−1999.355	−33018.786	−33024.283	−6506.555	−6487.148	−224025.433	−224049.114
	Min Δ(ln P)	−141.820	−142.271	−46.814	−47.045	−622.560	−622.292	−227.062	−227.516	−3371.980	−3361.094
	ln P(M) CI	±1.607	±1.318	±0.489	±0.680	±2.958	±5.619	±3.619	±5.268	±4.903	±5.084
	Δ(ln P)/Site	−0.373	−0.374	−0.593	−0.596	−0.548	−0.547	−0.316	−0.317	−1.686	−1.681

The rows are described in the main text. The mean $\Delta \ln P/\text{site} = -0.458$ (not including the synthetic data) and strongly implies preference for the relative-rate model.

data our method appears able to accurately estimate the rate even when multiple substitutions occur. In other words, over short times isoleucine will frequently substitute with leucine, but over long times a substitution to glutamine is highly informative as to the true underlying rate. As compared with the lower range, the middle range of substitution rates appear to have significantly less variance associated with them.

At greater than $r_i \approx 4.0$, Figure 1 shows that the sequences do become randomized with respect to each other, overwhelming even inter-class substitution events. Rather than estimate an excessively large rate, however, constraint (3) appears to bias the inferred rate downward. Thus, the model appears to be self-limiting with respect to high evolutionary rates without the a priori assumption of an exponential rate prior. Note that although the variance of high-rate parameters appears to be relatively small in the figure, the logarithmic scaling of the ordinate implies a larger variance than is visually evident.

2.3 Model comparison

For given fixed alignment and phylogenetic tree data D, both Maximum Likelihood (ML) and Bayesian estimations of the posterior rate distribution were performed. Alignments were initially computed with T-COFFEE (Notredame *et al.*, 2000) and then refined by inspection. Phylogenetic trees were inferred by PHYML (Guindon and Gascuel, 2003) using an optimized gamma model of rate heterogeneity. The WAG substitution matrix (Whelan and Goldman, 2001) was used throughout.

2.4 Protein families

Three proteins from two distinct families were studied to compare our relative-rate model to the more traditional independent-rate model. Specifically, Myc and Max, and two variants of p53 alignments were selected due to our familiarity with these families.

The Myc-Max-Mad network of basic-Helix-Loop-Helix (bHLH) transcription factor proteins is essential for control of cell growth, proliferation, differentiation and apoptosis. *Myc* is a well-established oncogene whose deregulated expression is responsible for a wide range of human cancers (Grandori *et al.*, 2000; Luscher, 2001). A comprehensive analysis of phylogeny and conservation in the bHLH-leucine-zipper (bHLHz) domain of a diverse set of Myc and Max homologs was performed by Atchley and Fernandes (2005) and is utilized herein.

In contrast, p53 belongs to the β-sandwich-domain family of DNA-binding transcription factors (Berardi *et al.*, 1999; Rudolph and Gergen, 2001) and is structurally independent of the bHLHz family. A detailed phylogenetic study of the p53 family has been presented by Fernandes and Atchley (2008). To mimic the situation where relatively few, closely related proteins are available for study, a subset of the p53 sequences, denoted p53R, was also analyzed.

2.5 Bayes factors

There is no straightforward procedure to contrast maximum likelihood and Bayesian models, but we and others have found that Bayes factors (Kass and Raftery, 1995) can be used to construct intuitively meaningful and statistically valid comparisons. Taking an approach similar to MRBAYES, we start with the independent site, gamma rate-prior model M_I and use Bayes factors to compare it to our relative-rate model M_R. Given model M_I, data D, a set of n independent-rate parameters $r = r_1, r_2, \ldots, r_n$, a shape parameter α, a likelihood model $P(D|r,\alpha,M_I)$ and prior distributions $P(r|\alpha,M_I)$ and $P(\alpha)$, Bayes' Theorem allows us to calculate

$$\begin{aligned} P(D|M_I) &= \iint P(D|r,\alpha,M_I)\cdot P(r|\alpha,M_I)\cdot P(\alpha)\,dr\,d\alpha \\ &= \int P(D|\alpha,M_I)\cdot P(\alpha)\,d\alpha \\ &\le \int P(D|\alpha,M_I)\cdot \delta(\alpha-\alpha_{\max})\,d\alpha \\ &= P(D|\alpha_{\max},M_I), \end{aligned}$$

where δ denotes Dirac's delta function and $\alpha_{\max}$ is the ML estimate of α. Since rates are independent under M_I, the first integration is standard and straightforward. The second integration over α acknowledges that we cannot know the 'correct' value of α exactly. Following standard Bayesian theory then, we draw it from some prior distribution $P(\alpha)$. Thus $P(D|M_I)$ will be maximal only if α is known precisely a priori and can only decrease as uncertainty about α increases. Thus, we use $P(D|\alpha_{\max},M_I)$ as a 'best case' conservative estimate of $P(D|M_I)$. The ratio of $P(D|M_I)$ to $P(D|M_R)$, known as the Bayes factor, indicates the relative weight of evidence supporting competing models H_I or H_R given the data.

Estimating $P(D|M_R)$ from the from the MCMC samples of the posterior likelihood is numerically challenging (Kass and Raftery, 1995). To estimate it, we utilized the stabilized harmonic mean estimator (Satagopan *et al.*, 2000) as provided by the MODEL_P program of BALIPHY (Redelings and Suchard, 2005; Suchard and Redelings, 2006). Comparative results are shown in Table 1. There, 'Max $\ln L$' denotes the maximum likelihood solution, '$\ln P(M)$' denotes the Bayesian probability of the relative-rate model,

'Max $\ln P$' denotes the maximum probability found for the relative-rate model during MCMC simulations and 'Min $\Delta\ln(P)$' denotes the minimum possible log-probability difference between the relative- and absolute-rate hypotheses; in other words, the smallest possible Bayes factor. More negative values indicate stronger support for the relative-rate model. Confidence intervals on $\ln P(M)$ are shown, along with the calculated $\Delta\ln P$/site. The latter is shown so that comparisons can be drawn between alignments of greatly different lengths.

For all examples studied, the minimum Bayes factor strongly supported the relative-rate model over site independence, with log-differences ranging from ≈ 47 to ≈ 3371. According to Jeffreys' scale (Jeffreys, 1961; Kass and Raftery, 1995) where differences of 2–10 are considered decisive, this represents overwhelming evidence in support of the relative model. Using long sampling times, the width of the $\ln P(M)$ confidence intervals were shortened to be insignificant compared to the magnitude of the Bayes factor. Since the magnitude of phylogenetic likelihoods tend to scale linearly with the number of alignment sites, the $\Delta\ln P$/site for each alignment was also calculated for each dataset. The values, ranging from -0.316 to -0.596 indicate that even short alignments of about 10 sites would overwhelmingly favor the relative-rate model.

The next most intriguing result displayed in Table 1 is the complete insensitivity of the model probability changes in the rate prior. Bayes factors are known to sometimes display extraordinary sensitivity to choice of prior (Kass and Greenhouse, 1989; Kass and Raftery, 1995). For the relative-rate model, however, no significant differences were detectable between the uniform and unit-exponential relative-rate prior: all differences were less than half the width of the model probability confidence interval. Again, we emphasize that the unit-exponential prior of the relative-rate model is *not* comparable to a unit-exponential independent-rate model. Although posterior probabilities are not significantly different between priors, a detailed comparison of the posterior densities would be required to recommend either as a suitable default.

2.6 Gamma shapes

Although the posterior rate distribution for the relative-rate model cannot be approximated by the independent gamma model, Figure 2 shows the distribution of 'best fit' gamma shape parameters across MCMC samples. Black circles denote the maximum likelihood shape parameter solution, while short horizontal bars indicate the mean shape parameter, along with quartiles and ranges. The ML shape parameter was always found to be outside the interquartile range of possible shapes. In the case of p53 and the synthetic datasets, the differences were substantial and indicate that the relative-rate posterior is significantly different than that implied by the independent-rate model.

Simply comparing ML shapes to 'best approximating' relative-rate shapes, however, fails to capture just how significantly different the posterior rate distributions are between $\overline{\mathbb{R}}^{n+}$ and $\mathbb{S}^n$. For instance, the best linear unbiased (BLU) estimator of central tendency in $\overline{\mathbb{R}}^{n+}$ is the arithmetic mean. For $\mathbb{S}^n$ the geometric mean (Pawlowsky-Glahn and Egozcue, 2002) is far more preferable. Furthermore, since the rates in $\mathbb{S}^n$ are by definition non-independent, the relative-rate posterior cannot be summarized by a scalar statistic.

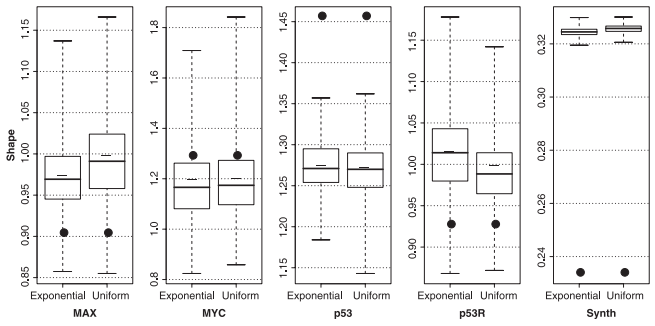

Fig. 2. Boxplots show the approximate distribution of estimated shape parameters from the MCMC integration; narrow internal lines show the mean shape estimate. Filled circles show the estimated shape parameter for the same system under maximum likelihood. In all cases the posterior shape distribution appears significantly different than that found by maximum likelihood.

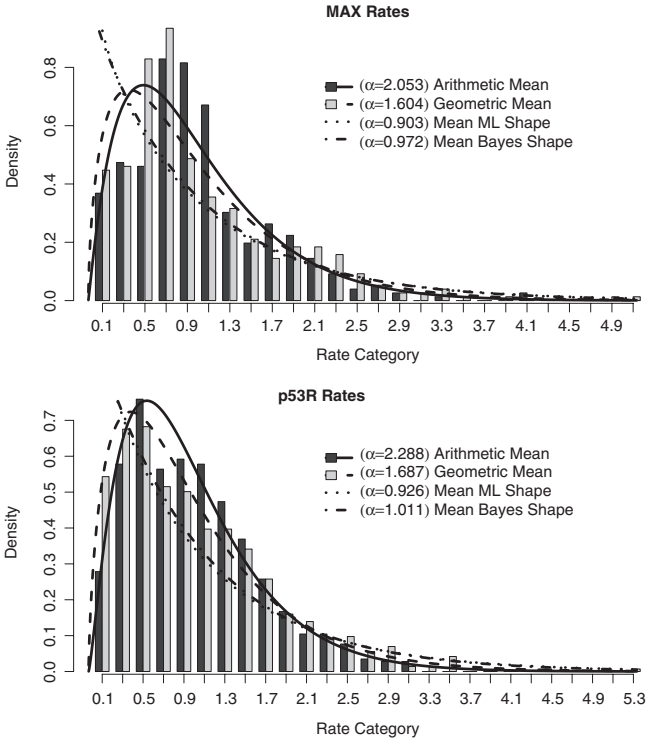

Fig. 3. The inferred distribution of rates for Max, showing the across-sample arithmetic and geometric means, as well as the best fit unit-gamma distribution shape parameter approximations between MCMC samples and of the final posterior mean.

Figure 3 illustrates just how different posterior rate estimates can be by comparing their best unit-gamma approximations. Shown are histograms of the posterior rate distribution for the relative-rate model taking either (a) the arithmetic or (b) geometric mean of all MCMC samples, along with best approximating shapes. The differences in distributions are striking, especially for the illustrated Max and p53R datasets. Also shown are (c) the best approximating

shapes for the maximum likelihood (independent-rate) model and (d) the mean relative-rate shape. In other words, the histograms show the mean rate of all MCMC samples, while the remaining curves show the best gamma distribution between MCMC samples (the mean of the sample shapes versus the shape of the mean rates). These figures support the idea that under the relative-rate model, the resultant posterior is inherently multivariate and cannot be correctly summarized by statistics of the marginals.

3 CONCLUSIONS

Current models in molecular evolution almost universally assume site independence to model rate heterogeneity. The unstated assumption is that as the number of sites increases, the independence model will become asymptotically more correct. Our results indicate that no simple relationship exists between the independent- and relative-rate models. The independent-rate model is conceptually simple although it requires considerable parameterization or foreknowledge of the rate prior and complicates branch length inference, requiring numerous regularization assumptions. The relative-rate model automatically encompasses a much greater class of rate prior without parameterization, results in better model fits, allows simultaneous branch-length inference, but is somewhat more computationally complex.

To see that no simple relationship exists between the models, consider the angle between the surface of $\mathbb{S}^n$ and the one of its bordering n-dimensional hyperplanes. Simple geometric arguments show that the radial angle between these subspaces is $\arcsin\left(\sqrt{(n-1)/n}\right)$. As $n\to\infty$ this angle approaches $\pi/2$, implying that as the dimension increases, the relative-rate model becomes orthogonal to any independent-rate model (minus one site). Although not a formal argument, this observation suggests that the relative-rate model cannot be easily approximated via an independent-rate model.

Comparisons with RATE4SITE (Pupko *et al.*, 2002) show similar marginal rate posteriors (data not shown). Differences are primarily observed when the marginal mean rate is either small or large. Thus, if it is known a priori that a given rate prior is appropriate for a given dataset, there may be no compelling reason to use the more accurate relative-rate model. It has been shown, however, that mixtures of gamma distributions often provide substantial model improvements in many situations (Mayrose *et al.*, 2005a, b). If little is known about the actual underlying rate distribution, then the relative-rate model is preferable since it does not require parameter estimation.

Again, we emphasize that although the *marginal* distributions computed by the independent- and relative-rate models often appeared *qualitatively* similar, Figures 2 and 3 emphasize that the intrinsic correlation present in the relative-rates model make between-model comparisons of marginal distributions virtually meaningless.

3.1 Posterior summarization

As shown in the figures, characterizing the rate posterior is not trivial. The Dirichlet distribution is often used as a summary distribution on the unit simplex, and can be readily fit to the posterior (Minka, 2003). However, Aitchison (1986) argues that the restrictive Dirichlet covariance structure make it surprisingly unsuitable for describing distributions on $\mathbb{S}^n$. Variants of the log-normal distribution is the preferred alternative. This alternative may hypothetically be used to study rate-heterogeneity covariance.

3.2 Hypervariability

Since substitution rates ≤ 4 substitutions per unit time appear to be resolvable if there is enough data, we postulate that *hypervariability* can be meaningfully defined for sites where the posterior rate is significantly and substantially greater than one. More investigation into the biological relevance of these sites is needed. In particular, preliminary observations indicate that some hypervariable sites are identifiable as homologous sites 'sandwiched' between conserved residues. However, other sites consist primarily of gaps, which are treated somewhat like an indeterminate amino acid, in standard likelihood calculations. Therefore, hypervariability may be biologically relevant in some situations, but not others.

3.3 ML formulation

Although presented in a Bayesian context, a maximum likelihood approach could be accommodated in an *unconstrained* optimization framework via composition analysis (Aitchison, 1986). Specifically, the isometric log-ratio (ILR) transformation (Egozcue *et al.*, 2003) can be used to construct a diffeomorphism between $\mathbb{S}^n$ and $\mathbb{R}^n$ under the standard Euclidian metric, with Jacobian $J\propto\left(\theta_1\theta_2\cdots\theta_n\right)^{-1}$. From an information-theoretic view, the ILR transformation uses this Jacobian as the invariant Haar measure on $\mathbb{S}^n$ and is equivalent to the use of Jaynes' prior (Jaynes, 1968). Thus rather than adaptively integrating over $\mathbb{S}^n$ as MCMC strives to do, it should be possible to find the most likely point $\theta\in\mathbb{S}^n$ via unconstrained optimization, and hence find $r=n\theta$.

4 METHODS

It has been suggested that MCMC sampling over $\mathbb{S}^n$ can be done by utilizing Dirichlet-distributed proposals (Larget and Simon, 1999). Our experience disagrees and shows that when n is large, sampling efficiency using Dirichlet proposals becomes intolerably low. The efficiency becomes particularly bad as $\theta\in\mathbb{S}^n$ approaches the boundary. Unfortunately, such approaches are common as they occur for all conserved sites.

To understand why the Dirichlet sampling is inefficient, suppose we are given the current Markov chain state as θ. A new state θ' is selected via

$$\theta'\sim\text{Dirichlet}\left(s\theta\right), \tag{4}$$

where s is a scalar scale factor. Under this parameterization, $\mathrm{E}\left[\theta'\right]=\theta$ and $\mathrm{Var}\left[\theta'\right]$ scale approximately as $1/s$. As θ approaches the simplex boundary, s must become very large to avoid inflating the sampling variance of θ'. A large value of s, however, implies that $\theta'-\theta$ must be small. Small MCMC sample differences imply long autocorrelation times, and hence intolerably inefficient sampling.

Instead, we developed a two-step MCMC sampling procedure with much higher sampling efficiency. If each marginal $\theta_i\overset{\text{iid}}{\sim}\Gamma\left(1,1\right)$, then $\theta/\sum_i\theta_i\sim\text{Dirichlet}\left(1,1,\ldots,1\right)$ is a standard result (Devroye, 1986). Therefore, given a current state θ, a new state θ' can be generated by the following procedure:

(1) For each component θ_i of θ, a new component θ_i' is sampled via MCMC such that the stationary distribution of θ_i' is unit-exponential.

(2) A secondary MCMC step is performed using $\theta_i'/\sum_i\theta_i'$ and the phylogenetic likelihood function.

(3) Repeat, using n individual θ_i parameters to hold the 'state' of the $n-1$ relative rates.

Thus the proposal function itself is first sampled via MCMC, and the resulting point is used to sample the relevant posterior. The procedure works because the acceptance or rejection of a given step is always, by definition, independent of the previous state. Furthermore, the sum of the state variables is statistically independent of each individual (Devroye, 1986). The efficiency of the algorithm is quite high as the exponential scaling of the marginals ensures that the new sample scales optimally along each dimension of the simplex.

ACKNOWLEDGEMENTS

A. D. F. would like to thank Lindi M. Wahl and Gregory B. Gloor for funding and mentorship. Data processing and analysis was done with R (R Development Core Team, 2008).

Funding: National Institutes of Health (GM45344); North Carolina State University; the Alexander von Humboldt Stiftung; the Canadian Institutes of Health Research and the Natural Sciences and Engineering Research Council of Canada.

Conflict of Interest: none declared.

REFERENCES

Aitchison,J. (1986) *The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability*. Chapman and Hall, New York, London.

Atchley,W.R. and Fernandes,A.D. (2005) Sequence signatures and the probabilistic identification of proteins in the myc-max-mad network. *Proc. Natl Acad. Sci. USA*, **102**, 6401–6406.

Berardi,M.J. *et al*. (1999) The Ig fold of the core binding factor αRunt domain is a member of a family of structurally and functionally related Ig-fold DNA-binding domains. *Structure*, **7**, 1247–1256.

Berger,J. (2006) The case for objective bayesian analysis. *Bayesian Anal.*, **1**, 385–402.

Bernardo,J.M. and Ramon,J.M. (1998) An introduction to bayesian reference analysis: inference on the ratio of multinomial parameters. *J. R. Stat. Soc. D*, **47**, 101–135.

Bernardo,J.M. and Smith,A. (1994) *Bayesian Theory*. John Wiley and Sons, New York.

Corbin,K. and Uzzell,T. (1970) Natural selection and mutation rates in mammals. *Am. Nat.*, **104**, 37–53.

Devroye,L. (1986) Non-uniform random variate generation. Available at http://cg.scs.carleton.ca/~luc/rnbookindex.html (last accessed, August 11, 2008).

Egozcue,J.J. *et al*. (2003) Isometric logratio transformations for compositional data analysis. *Math. Geol.*, **35**, 279–300.

Felsenstein,J. (2001) Taking variation of evolutionary rates between sites into account in inferring phylogenies. *J. Mol. Evol.*, **53**, 447.

Felsenstein,J. (2004) *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA.

Fernandes,A.D. and Atchley,W.R. (2006) Gaussian quadrature formulae for arbitrary positive measures. *Evol. Bioinform.*, **2**, 261–269.

Fernandes,A.D. and Atchley,W.R. (2008) Biochemical and functional evidence of p53 homology is inconsistent with molecular phylogenetics for distant sequences. *J. Mol. Evol.* **67**, 51–67.

Grandori,C. *et al*. (2000) The myc/max/mad network and the transcriptional control of cell behavior. *Annu. Rev. Cell Dev. Biol.*, **16**, 653–699.

Gu,X. *et al*. (1995) Maximum-likelihood-estimation of the heterogeneity of substitution rate among nucleotide sites. *Mol. Biol. Evol.*, **12**, 546–557.

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

Jaynes,E.T. (1968) Prior probabilities. *IEEE T. Syst. Sci. Cyb.*, **4**, 227–241.

Jeffreys,H. (1946) An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond. A*, **186**, 453–461.

Jeffreys,H. (1961) *Theory of Probability*. 3rd edn. Clarendon Press, Oxford.

Kass,R. and Greenhouse,J. (1989) Comments on "investigating therapies of potentially great benefit: ECMO". *Stat. Sci.*, **4**, 310–317.

Kass,R. and Raftery,A. (1995) Bayes factors. *J. Am. Stat. Assoc.*, **90**, 773–795.

Kass,R.E. and Wasserman,L. (1996) The selection of prior distributions by formal rules. *J. Am. Stat. Assoc.*, **91**, 1343–1370.

Kelly,C. and Rice,J. (1996) Modeling nucleotide evolution: a heterogeneous rate analysis. *Math. Biosci.*, **133**, 85–109.

Kimura,M. (1983) The neutral theory of molecular evolution. In Nei,M. and Koehn,R. (eds), *Evolution of Genes and Proteins*, Sinauer Associates, Sunderland, MA, pp. 208–233.

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

Luscher,B. (2001) Function and regulation of the transcription factors of the mye/max/mad network. *Gene*, **277**, 1–14.

Mayrose,I. *et al*. (2004) Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol. Biol. Evol.*, **21**, 1781–1791.

Mayrose,I. *et al*. (2005a) A gamma mixture model better accounts for among site rate heterogeneity. *Bioinformatics*, **21**, 151–158.

Mayrose,I. *et al*. (2005b) Site-specific evolutionary rate inference: taking phylogenetic uncertainty into account. *J. Mol. Evol.*, **60**, 345–353.

Meyer,S. and von Haeseler,A. (2003) Identifying site-specific substitution rates. *Mol. Biol. Evol.*, **20**, 182–189.

Minka,T.P. (2003) Estimating a dirichlet distribution. *Technical report*. Microsoft Research. Available at http://research.microsoft.com/~minka/papers/dirichlet/ (last accessed, August 11, 2008)

Nei,M. *et al*. (1976) Infinite allele model with varying mutation rate. *Proc. Natl Acad. Sci. USA*, **73**, 4164–4168.

Notredame,C., *et al*. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.

Pawlowsky-Glahn,V. and Egozcue,J. (2002) BLU estimators and compositional data. *Math. Geol.*, **34**, 259–274.

Pond,S.L.K. and Frost,S.D.W. (2005) A simple hierarchical approach to modeling distributions of substitution rates. *Mol. Biol. Evol.*, **22**, 223–234.

Pupko,T. *et al*. (2002) Rate4site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, **18**(Suppl. 1), S71–S77.

R Development Core Team (2008) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Available at http://www.R-project.org, Vienna, Austria (last accessed, August 11, 2008).

Redelings,B.D. and Suchard,M.A. (2005) Joint Bayesian estimation of alignment and phylogeny. *Syst. Biol.*, **54**, 401–418.

Ronquist,F. and Huelsenbeck,J.P. (2003) Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.

Rudolph,M.J. and Gergen,J.P. (2001) DNA-binding by Ig-fold proteins. *Nat. Struct. Mol. Biol.*, **8**, 384–386.

Satagopan,J. *et al*. (2000) Easy estimation of normalizing constants and Bayes factors from posterior simulation: stabilizing the harmonic mean estimator. *Technical Report 382*. University of Washington Available at http://www.stat.washington.edu/research/reports/2000/tr382.pdf (last accessed, August 11, 2008).

Suchard,M.A. and Redelings,B.D. (2006) BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics*, **22**, 2047–2048.

Swofford,D. *et al*. (1996) Phylogenetic inference. In Hillis,D. *et al*. (eds) *Molecular Systematics*. 2nd edn. Sinauer, Sunderland, Massachusetts, pp. 407–514.

Syversveen,A.R. (1998) Noninformative Bayesian priors. interpretation and problems with construction and applications. *Technical Report 3/98*. Institutt for Matematiske Fag. Available at http://www.math.ntnu.no/preprint/statistics/1998/S3-1998.ps (last accessed, August 11, 2008).

Uzzell,T. and Corbin,K. (1971) Fitting discrete probability distributions to evolutionary events. *Science*, **172**, 1089–1096.

Wasserman,L. (2000) Asymptotic inference for mixture models using data-dependent priors. *J. R. Stat. Soc. B*, **62**, 159–180.

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

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. (1996) Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.*, **11**, 367–372.

Yang,Z.H. and Kumar,S. (1996) Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. *Mol. Biol. Evol.*, **13**, 650–659.