BIOINFORMATICS

Vol. 21 Suppl. 2 2005, pages ii151–ii158
doi:10.1093/bioinformatics/bti1125

Phylogenetics

A Gamma mixture model better accounts for among site rate heterogeneity

Itay Mayrose[1], Nir Friedman[2] and Tal Pupko[1,*]

[1]Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel and [2]School of Computer Science and Engineering, Hebrew University, Jerusalem 91904, Israel

ABSTRACT

Motivation: Variation of substitution rates across nucleotide and amino acid sites has long been recognized as a characteristic of molecular sequence evolution. Evolutionary models that account for this rate heterogeneity usually use a gamma density function to model the rate distribution across sites. This density function, however, may not fit real datasets, especially when there is a multimodal distribution of rates. Here, we present a novel evolutionary model based on a mixture of gamma density functions. This model better describes the among-site rate variation characteristic of molecular sequence evolution. The use of this model may improve the accuracy of various phylogenetic methods, such as reconstructing phylogenetic trees, dating divergence events, inferring ancestral sequences and detecting conserved sites in proteins.

Results: Using diverse sets of protein sequences we show that the gamma mixture model better describes the stochastic process underlying protein evolution. We show that the proposed gamma mixture model fits protein datasets significantly better than the single-gamma model in 9 out of 10 datasets tested. We further show that using the gamma mixture model improves the accuracy of model-based prediction of conserved residues in proteins.

Availability: C++ source codes are available from the authors upon request.

Contact: talp@post.tau.ac.il

1 INTRODUCTION

Probabilistic models of evolution are considered to be the-state-of-art methods for phylogeny inference. Likelihood methods calculate the probability of observing the data conditioned on the hypothesis, as specified by the phylogenetic tree and the probabilistic evolutionary model. The maximum-likelihood (ML) principle is then invoked to choose the hypothesis that yields the highest likelihood of the observed sequences. This paradigm allows for statistically robust parameter estimation and vigorous testing of evolutionary hypotheses (Whelan *et al*., 2001).

A basic dilemma when using probabilistic models within the ML paradigm is controlling the expressiveness of the model. Models with too many parameters might overfit the observations. However, models with too few parameters may be unrealistic, resulting in erroneous conclusions. A classical example of oversimplification is the assumption of equal evolutionary rates at all sites of a protein (Felsenstein, 2001). Nevertheless, in proteins the rates of evolution vary due to different selective constraints that are acting on different sites. Indeed, a vital advance in the reconstruction of evolutionary trees has been the consideration of heterogeneity of evolutionary rates among sequence sites (reviewed in Swofford *et al*., 1996; Yang, 1996). Accordingly, the rate at each site is modeled as a random variable drawn from a specified prior distribution. By far, the most commonly chosen distribution for modeling rate variation across sites is the gamma distribution.

Assuming a gamma prior over the rate distribution is mathematically convenient and requires the fitting of a single additional parameter. However, as noted by Felsenstein (2001), there is no reason to believe that the rate across sites is gamma distributed. Biological insight suggests that rate distributions are not always unimodal. For example, a protein may be composed of several domains, each having its own rate distribution. As shown in Figure 1A and B, the unimodal gamma distribution poorly fits the observed rate distribution of the adenylate kinase protein family.

In a step toward a more realistic evolutionary model we propose a gamma mixture model, which generalizes the traditional single-gamma distribution model. This model assumes the existence of K gamma distributions (components), each characterized by its own set of parameters. The model further defines the a priori probability for each gamma component. The adjustable parameters of the model are optimized using an expectation–maximization (EM) algorithm. The resulting model can accommodate a multimodal rate distribution, where the number of modes depends on the number of gamma components and how different each component is from another (Fig. 1C).

The outline of the paper is as follows. Section 2 presents and formulates the gamma mixture model of among-site rate variation. In Section 3 we develop an efficient EM algorithm for estimating the model parameters. In Section 4 we apply the gamma mixture model to a wide range of datasets. We show that the mixture model significantly outperforms the traditional single-gamma model. In Section 5 we study the relationship between the data size and the number of gamma components that the data support. In Section 6 we demonstrate that by using the gamma mixture model more accurate predictions of the conserved residues within a protein are obtained. We conclude with a discussion (Section 7).

2 THEORY

2.1 Among-site rate variation

A phylogenetic tree $T = (\tau, t)$, is defined by its tree topology τ and associated branch lengths t. The branch lengths of the phylogenetic

*To whom correspondence should be addressed.

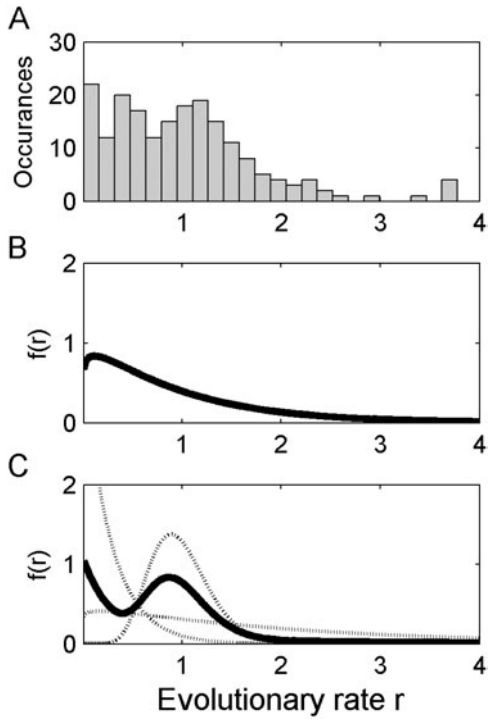

Fig. 1. An example of the rate distribution inferred for the adenylate kinase protein family (first row in Table 1). (**A**) The rate distribution as inferred using ML with the Rate4Site program (Pupko *et al*., 2002). (**B**) The fitted rate distribution inferred using a model that assumes one-gamma component is unimodal. (**C**) The fitted rate distribution inferred using a model that assumes three gamma components is bimodal. The individual gamma components are shown by dashed lines. The method used for estimating the model parameters is described in Section 3.

tree represent the average evolutionary rate across all sites. The substitution model describes how characters (amino acids, nucleotides or codons) evolve on the tree. A phylogenetic tree and a substitution model induce probabilities over assignments of characters to the leaves of the tree (see Felsenstein, 2004 for a detailed description). Let D denote a dataset that consists of aligned sequences of current day taxa corresponding to the leaves of the tree. In the standard ML models, we view each column of the alignment as evolving independently. Thus, if we denote by D_i the data at site i, then $P(D_i|T)$ is the probability of the i-th column of the alignment given the tree. Computing this probability requires summing over all possible character assignments to internal nodes of the tree (ancestral states). This computation can be done efficiently using Felsenstein's (1981) post-order tree traversal algorithm.

The standard ML model assumes that each site evolves at the same rate. However, biology suggests that some sites are more conserved and undergo fewer substitutions, whereas other sites are less conserved and undergo more substitutions. Since longer branches imply more substitutions, we can model such differences by shrinking or expanding the branch lengths. The site-specific rate, r_i, indicates how fast site i evolves compared with the average rate across all sites. For example, a rate of 2 indicates a site that evolves twice as fast as the average. Thus, site-specific rates are not absolute evolutionary rates that require knowledge of divergence times, but instead represent a comparative quantity. We define $T \times r$ to be the tree T with all branch lengths multiplied by rate r.

Since we do not know the actual value of r_i, we need to consider all possible values when computing the probability of D_i. Thus, the likelihood of site i is defined as

$$L^{(i)} = \int_0^\infty P(D_i|T \times r_i) g(r_i : \theta) dr_i, \tag{1}$$

and the likelihood of the complete alignment is the product

$$L = \prod_i L^{(i)}, \tag{2}$$

where $g(r : \theta)$ is a prior density function over the rates which is governed by parameters θ. Choosing θ determines how rate variation is modeled. Here we focus on the key question of how to choose the prior $g(r : \theta)$ and its effect.

2.2 The gamma distribution

The most commonly used prior distribution over evolutionary rates is the gamma distribution (Swofford *et al*., 1996; Yang, 1996). This distribution has two parameters; a shape parameter, α, and a scale parameter, β. A variable R is gamma distributed, denoted $R \sim \Gamma(\alpha, \beta)$, if its density function is

$$g(r \cdot \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \mathrm{e}^{-\beta r} r^{\alpha-1}. \tag{3}$$

The mean of the gamma distribution is α/β. Standard applications of the gamma distribution prior require that the mean of the prior is 1 (otherwise, we can rescale the original tree). This implies that $\beta = \alpha$, leaving one free parameter. The shape of the gamma distribution is determined by the α parameter, which is indicative of rate variation. When $\alpha > 1$ the distribution is bell-shaped, suggesting little rate heterogeneity. In the case of $\alpha < 1$, the distribution is highly skewed and is L-shaped, which indicates high levels of rate variation. This flexibility makes the distribution suitable for accommodating different levels of rate variation in different datasets.

2.3 A mixture of gamma

We suggest modeling the prior density over evolutionary rates by a mixture of gamma distributions. This assumes that the rates are pulled from a few possible gamma components, each with its own α and β parameters. Let K be the number of such gamma components, with α_k and β_k being the parameters of the k-th component. We denote by γ_k the prior probability that a specific rate was drawn from the k-th distribution. We now have $\theta = \{\langle \alpha_k, \beta_k, \gamma_k \rangle : k = 1, \ldots, K\}$, with $\sum_{k=1}^K \gamma_k = 1$. A variable R is distributed with a mixture of gamma if

$$gm(r : \theta) = \sum_{k=1}^K \gamma_k g(r : \alpha_k, \beta_k), \tag{4}$$

where $g(r : \alpha_k, \beta_k)$ is the gamma distribution of the k-th component. Such a mixture provides additional flexibility in specifying the prior distribution. Moreover, by choosing K we can consider a range of distributions with growing expressiveness with corresponding increase in the number of parameters. Since we restrict the expectation of the mixture to equal 1, the number of free parameters for K components is $3K - 2$.

2.4 Likelihood computation

An important technical issue is how to compute the likelihood $L^{(i)}$ with its integration. The standard solution is to approximate the integral by a weighted sum over a set of discrete rates. Thus, we approximate the integral in Equation (1) by a sum

$$L^{(i)} \cong \sum_{j=1}^{S} P(D_i|T \times \tilde{r}_j) w(\tilde{r}_j), \tag{5}$$

where $(\tilde{r}_1, \ldots, \tilde{r}_S)$ are representative rates and $(w(\tilde{r}_1), \ldots, w(\tilde{r}_S))$ are the corresponding weights.

For the case of a single gamma prior a common choice is based on Yang's (1994) quantile approximation, where the weights of representative rates are identical. Here, we apply a more accurate approximation based on the generalized Laguerre quadrature method as suggested by Felsenstein (2001). A detailed description of the Laguerre quadrature method is given in the Appendix.

We can also use Laguerre quadrature for approximating $L^{(i)}$ with a mixture of gamma distribution by finding the representative weights of each component, and then use the approximation

$$L^{(i)} \cong \sum_{k=1}^{K} \gamma_k \sum_{j=1}^{S} P(D_i|T \times \tilde{r}_j^k) w(\tilde{r}_j^k), \tag{6}$$

where $\tilde{r}_j^k$ is the rate of the j-th discrete rate category in the k-th component. Since now each gamma component is approximated by S discrete categories the total number of categories is $K \times S$.

3 AN EM ALGORITHM FOR OPTIMIZING THE GAMMA MIXTURE PARAMETERS

When the parameters of the mixture model are known, computing the likelihood is simple. However, the parameters of the model are usually unknown and have to be estimated for each dataset. In this section we address this multidimensional maximization problem. We adopt an ML approach and aim to maximize $P(D|T, \theta)$. Our approach for maximizing the likelihood is to use an EM procedure (Dempster *et al*., 1977) that is described below. For succinctness, we assume throughout this discussion that the phylogenetic tree T is fixed, and we do not explicitly refer to it in the equations.

To develop the EM procedure, it is useful to introduce a random variable H_i, which explicitly denotes the mixture component at site i. Our model can thus be cast as a graphical probabilistic model (Buntine, 1994) as shown in Figure 2. We want to estimate the vector parameters α, β and γ. We start by reviewing how to estimate these parameters from complete data where we observe H_i and R_i for each site, and then consider the more challenging case where we do not observe them.

Consider the complete data case where M is the sequence length. The distribution $P(H_i : \gamma)$ is a multinomial distribution. The sufficient statistics for this distribution are

$$M_k = \sum_{i=1}^{M} 1\{H_i = k\}, \tag{7}$$

where $1\{\ \}$ is the indicator function. Given the vector $\{M_k : k = 1, \ldots, K\}$, the ML estimate for γ is simply M_k/M.

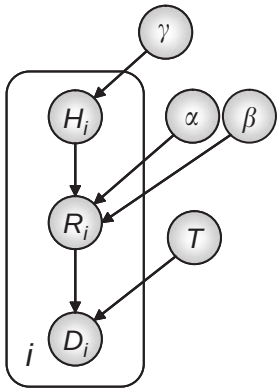

Fig. 2. A graphical plate model representation of the mixture of gamma distribution model. Site-specific variables are shown within the plate. The variables outside the plate are parameters shared among all positions. The variable H_i denotes the mixture component, R_i the rate and D_i the observed characters for the i-th site.

The distribution $P(R_i|H_i = k : \alpha_k, \beta_k)$ is a gamma distribution. The sufficient statistics for this distribution are the sum of R and the sum of $\ln R$ in those sites where $H_i = k$. Thus,

$$A_k = \sum_{i=1}^{M} 1\{H_i = k\} R_i, \tag{8}$$

$$B_k = \sum_{i=1}^{M} 1\{H_i = k\} \ln R_i. \tag{9}$$

Finding the ML estimates of α_k and β_k requires solving a pair of equations

$$B_k = \frac{M_k}{A_k} \alpha_k \tag{10}$$

and

$$0 = \ln M_k + \ln \alpha_k - \ln A_k + \frac{B_k}{M_k} - \phi(\alpha_k) \tag{11}$$

where $\phi(\alpha) = \Gamma'(\alpha)/\Gamma(\alpha)$ is the digamma function that can be approximated by a series expansion (Abramowitz and Stegun, 1972). The second equation cannot be solved analytically, but can be solved numerically by a line search.

To conclude, given complete data, we can accumulate the sufficient statistics $\{M_k, A_k,$ and $B_k, k = 1, \ldots, K\}$ and then find the ML estimates of α, β, and γ.

Unfortunately, we do not observe the variables H_i and R_i for different sites. Thus, we need to learn from incomplete data. The EM algorithm for learning with incomplete data uses the 'plug in' estimators of the complete data case as a sub-procedure. The algorithm starts with an initial estimate of the unknown parameters and iteratively improves them. Let θ^t denote the parameters after the t-th iteration. Starting with a guess, θ^0, the EM algorithm iterates between the following two steps:

E-step. Compute the expected sufficient statistics given the data and the parameters θ^t. This requires computing $E[M_k|D, \theta^t]$, $E[A_k|D, \theta^t]$ and $E[B_k|D, \theta^t]$ for each k (see Equations (15–18) below).

M-step. Given these expected sufficient statistics, set θ^{t+1} to be the ML estimate as though these statistics were obtained from complete data. In our case, this implies solving

$$0 = \ln E[M_k|D,\theta^t] + \ln \alpha_k^{t+1} - \ln E[A_k|D,\theta^t] + \frac{E[B_k|D,\theta^t]}{E[M_k|D,\theta^t]} - \phi(\alpha_k^{t+1}); \quad (12)$$

$$\beta_k^{t+1} = \frac{E[M_k|D,\theta^t]}{E[A_k|D,\theta^t]}\alpha_k^{t+1}; \quad (13)$$

$$\gamma_k^{t+1} = \frac{E[M_k|D,\theta^t]}{M}. \quad (14)$$

Each EM iteration is monotonically non-decreasing in terms of the data likelihood (Dempster *et al*., 1977). The algorithm converges (i.e. the likelihood does not increase) only at stationary points of the likelihood function. Thus, EM will converge to local maxima of the likelihood surface. The quality of the convergence point depends on the starting guess and the properties of the specific problem.

The key computational step in performing EM is computing the expected sufficient statistics given the parameters θ^t. Using linearity of expectations, we can rewrite these as

$$E[M_k|D,\theta^t] = \sum_{i=1}^{M} P(H_i = k|D_i,\theta^t)$$

$$E[A_k|D,\theta^t] = \sum_{i=1}^{M} P(H_i = k|D_i,\theta^t)E[R_i|H_i = k, D_i,\theta^t] \quad (15)$$

$$E[B_k|D,\theta^t] = \sum_{i=1}^{M} P(H_i = k|D_i,\theta^t)E[\ln R_i|H_i = k, D_i,\theta^t].$$

These computations require the following terms:

$$P(H_i = k|D_i,\theta^t) = \frac{P(H_i = k, D_i|\theta^t)}{P(D_i|\theta^t)}; \quad (16)$$

$$P(H_i = k, D_i|\theta^t) = \int_0^\infty P(r_i, H_i = k, D_i|\theta^t)dr_i = \gamma_k^t \int_0^\infty P(D_i|r_i)g(r_i : \alpha_k^t, \beta_k^t)dr_i; \quad (17)$$

$$E[R|H_i = k, D_i,\theta^t] = \int_0^\infty r_i P(r_i|H_i = k, D_i,\theta^t)dr_i = \frac{\int_0^\infty r_i P(D_i|r_i)g(r_i : \alpha_k^t, \beta_k^t)dr_i}{P(D_i|H_i = k,\theta^t)}. \quad (18)$$

$E[\ln R_i|H_i = k, D_i,\theta^t]$ can be similarly obtained. These terms require integration of terms of the form $f(r_i)g(r_i : \alpha_k, \beta_k)$. Again, we use Laguerre quadrature to approximate these integrals (Appendix).

4 ANALYSIS OF EXAMPLED DATASETS

4.1 Datasets

To test the utility of the proposed model we selected 10 datasets from the HSSP database (Sander and Schneider, 1993). We refer to each dataset by its protein data bank (PDB; Sussman *et al*., 1998) identifier. These datasets encompass a broad range of organisms (from bacteria to mammals) and biological processes, such as cell growth (3adk), metabolism (3pgk, 1rhd, 1ppl, 1rnd, 4mt2 and 3dfr), apoptosis (1bxl), ion transport (1bl8) and signal transduction (1a6q). Sequences with many gapped positions were manually removed. To avoid prohibitive computation time, we limited the number of sequences such that the 50 most divergent sequences in each dataset were selected. Gapped positions were treated as missing characters.

For each dataset the tree topology was constructed according to the neighbor-joining (NJ) algorithm (Saitou and Nei, 1987) with pairwise distances estimated by ML with a homogenous rate model. Branch lengths in the resulting tree and the gamma mixture parameters were then optimized iteratively until convergence of the likelihood function. The choice of the number of discrete rate categories can slightly influence the resulting likelihood score. In order to equally compare results of models with different number of components, we chose the total number of categories in each model to be 36. For example, a model with 2 components had 18 discrete categories in each component. We note that our results were essentially the same when the number of categories per component was identical for all models, although this second discretization scheme slightly favored models with more components (data not shown).

4.2 The evolutionary model

All analyses conducted in this study used the JTT model of amino acid replacements (Jones *et al*., 1992). However, incorporating the gamma mixture model into any desired nucleotide or amino acid substitution model is an easy extension.

4.3 Model comparisons

In this study we considered models with 1, 2 and 3 components, denoted by M1, M2 and M3, respectively. We used the likelihood ratio test (LRT) to test whether models with larger number of components fit a particular dataset significantly better than a model with fewer components. Table 1 contains maximum log-likelihood estimates obtained when analyzing each of the 10 datasets with models M1, M2 and M3. Adding one-gamma component requires three more free parameters (α and β, the distribution parameters and γ, the component probability) and is statistically justified if the log-likelihood improvement is >3.95 ($P < 0.05$; χ^2 with 3 degrees of freedom). In 9 of 10 datasets examined (Table 1) M2 gives a significantly better fit to the data than the commonly used M1 model. The only exception is the 4mt2 dataset. One explanation to this exception is that the sequence length of the proteins in this dataset is only 62 amino acids long (see Section 5). In contrast to the significant difference between M1 and M2, the addition of a third gamma component (M3) is statistically unjustified in all but the 3pgk dataset, i.e. the dataset with the longest sequence length. We note that although the models are nested, the use of the LRT may not be justified because of boundary problems (Anisimova *et al*., 2001). In addition, the LRT does not account for the size of the data. We thus also used a second order Akaike Information Criterion (AIC_c) (Burnham and Anderson, 2002), defined as $AIC_c = -2LL + 2pM/(M - p - 1)$, where LL is the log-likelihood, p the number of free parameters and M is sequence length. In all datasets examined significant LRT results were also supported by better AIC_c scores. The exact distribution of the LRT statistics can be approximated using parametric bootstrapping (Susko *et al*., 2003). However, this approach is computationally intensive and was not applied here.

Table 1. Maximum log-likelihood (LL) estimates for the analysis of 10 datasets under the M1, M2, M3 and G + I models

Dataset	NS[a]	SL[b]	M1	M2[c]	M3[c]	G + I
3adk	50	194	−16429.3	**−16406.5**	−16406.1	−16423.1
3pgk	50	415	−31380	**−31364.3**	**−31359.7**	−31376
1rhd	50	293	−17583.3	**−17572.5**	−17571.3	−17579
1ppl	50	323	−26402	**−26389.7**	−26387	−26397
1rnd	50	124	−8388.9	**−8378.5**	−8377.5	−8382.5
4mt2	50	61	−1000.4	−997.8	−997.6	−1000.3
3dfr	50	163	−12831.9	**−12820.3**	−12818.5	−12829.4
1bxl	36	181	−3155.2	**−3141.3**	−3140.5	−3155.2
1bl8	50	97	−8352.6	**−8343.7**	−8340.2	−8347.7
1a6q	50	363	−27470.7	**−27461.2**	−27459.1	−27468.1

[a]Number of sequences.
[b]Sequence length.
[c]Observed LL values are shown in bold type if the LRT between M2 (M3) and M1 (M2) is significant ($P < 0.05$).

We have also compared the gamma mixture model to the Gamma+ Invariant (G + I) site model (Gu *et al*., 1995). This model may represent a variant of M2 in which the second gamma component is constrained to invariable rates only. As expected, the G + I model gave intermediate log-likelihoods between M1 and M2 (Table 1). Nevertheless, for all datasets examined, the AIC_c scores of M2 were better than that of G+I. For some datasets (3adk, 3pgk and 1bxl) the AIC_c difference was >20, whereas in the 4mt2 dataset the superiority of M2 was only marginal. For the datasets tested, it seems that M2 superiority over G + I was most pronounced in those cases where the expectation of the component with the lower average rate (the 'slower' gamma component) highly deviated from zero. For example, in all cases where the AIC_c difference was >20 the expectation of the 'slower' gamma component was >0.4, resulting in a relatively large log-likelihood difference between M2 and G + I.

5 THE INFLUENCE OF DATA SIZE

We tested the hypothesis that adding more components to the model will become more justified as the sequence length increases. Two datasets were analyzed (3adk and 3pgk). In each dataset we randomly selected *l* sites to produce a new alignment that is a subset of the original multiple sequence alignment (MSA). We then optimized the gamma mixture parameters and branch lengths (only two rounds of optimization iterations were performed since each such iteration is computationally intensive). For each sequence length, 10 independent runs were conducted. As shown in Figure 3A (for the 3adk dataset) the average difference between models M2 and M1 increases as the sequence length increases. The same trend, though to a lesser extent, is also apparent when comparing M3 and M2. Similar results were obtained for the 3pgk dataset (data not shown).

We further analyzed whether the inclusion of more sequences in the MSA contributes to the fit of more complex models. We analyzed this hypothesis with the 3adk and 3pgk datasets. *s* random sequences were sampled from each dataset to produce a new MSA. For each value of *s* (ranged from 10 to 80), 10 independent runs were conducted. Our results indicate that the average difference between M2 and M1 increases as the number of sequences increases (Fig. 3B). Again, this is also true for the M3 versus M2 comparison to a lesser extent. Similar results were obtained for the 3pgk dataset (data not shown). To conclude, when more information is available (number of sequences and sequence length) models that account for a complex distribution of rates better explain the data.

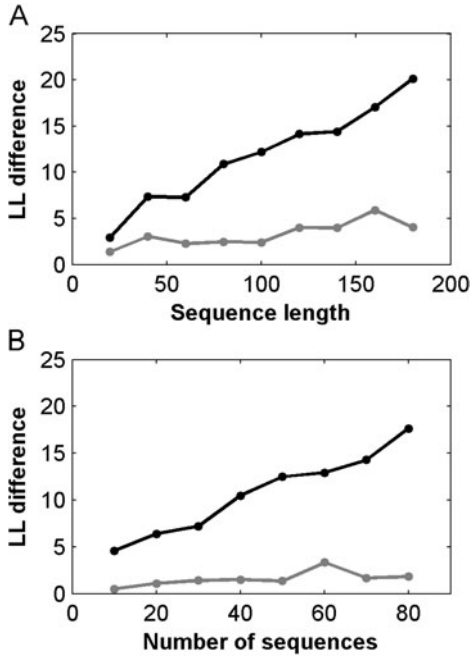

Fig. 3. Log-likelihood (LL) difference obtained for the 3adk dataset as a function of (**A**) sequence length and (**B**) number of sequences. The LL difference is the average score obtained over 10 independent runs. The differences between models M2 and M1, and between M3 and M2 are shown in black and grey lines, respectively.

6 BAYESIAN ESTIMATION OF SITE-SPECIFIC EVOLUTIONARY RATES USING THE GAMMA MIXTURE MODEL

Our results above indicated that the mixture model can statistically explain sequence variability better than the traditional single-gamma

model. We now investigate whether the use of our suggested model can also improve methods that aim to estimate the rate at which a site evolves as a means for inferring conserved and variable sites of a protein.

Within the Bayesian framework, the posterior probability of any given rate, r, is obtained from the likelihood function and the prior distribution. Given an estimated phylogeny and a discrete gamma mixture prior distribution (for which $w(\tilde{r}_j^k)$ is the probability of rate category j in component k), this probability is then

$$P(R_i = \tilde{r}_j^k | D_i, T, \theta) \cong \frac{P(D_i|\tilde{r}_j^k, T)\gamma_k w(\tilde{r}_j^k)}{\sum_{k=1}^{K}\sum_{j=1}^{S} P(D_i|\tilde{r}_j^k, T)\gamma_k w(\tilde{r}_j^k)}. \quad (19)$$

Our site-specific rate estimate is defined as the expectation of r over its posterior rate distribution

$$E(R_i|D_i, T, \theta) \cong \sum_{k=1}^{K}\sum_{j=1}^{S} P(R_i = \tilde{r}_j^k | D_i, T, \theta)\tilde{r}_j^k. \quad (20)$$

6.1 Simulations setup

We have previously shown that an empirical Bayesian site-specific rate inference method is superior to an ML-based approach (Mayrose *et al*., 2004). Here, simulations were used in order to test whether a prior that assumes a mixture of gamma components (models M2 and M3) can improve the accuracy of site-specific rate estimates over a model with a single component (M1). We simulate a given site with a specific 'true' rate. An MSA is thus generated based on a vector of true rates. Subsequently, a rate for each column is inferred using either the M1, M2 or M3 model. The closer the inferred rates to the true rates are, the better the inference is. For the simulations, one must determine the true rate in each site. In order to obtain true rates that are biologically relevant, characteristic rates were computed based on two empirical datasets: 3adk with 172 sequences and 194 sites, and 3pgk with 150 sequences and 415 sites. For each dataset, ML was used to estimate the site specific rates. Thus, no prior was assumed when constructing the empirical rate distribution. (Inferring the rates using a gamma mixture with a specified number of components would bias the results toward this specific distribution.) True rates were scaled such that the average was set to 1. For estimating the ML rates, the 3adk and 3pgk phylogenetic trees were reconstructed using NJ (Saitou and Nei, 1987).

For each dataset and for each number of sequences tested, a total of 20 identical and independent simulation runs were conducted. The accuracy of inference was analyzed by the mean relative absolute deviations (MRAD) distance between the simulated and estimated rates:

$$\text{MRAD} = \frac{1}{M}\sum_{i=1}^{M}\frac{|\text{estimated } r_i - \text{true } r_i|}{\text{true } r_i}, \quad (21)$$

where M is the sequence length. The division of each absolute deviation by the true rate compensates for the larger variance in large rates and the smaller variance in low rates.

A two-sided Wilcoxon non-parametric test between two dependent samples (Sokal and Rohlf, 1981) was then performed in order to determine whether the difference in accuracy obtained from models with a different number of gamma components is statistically significant. A non-parametric test was used to eliminate the assumption that the MRAD measures are normally distributed, although similar results were obtained with a parametric t-test (data not shown). For each empirical dataset, the influence of the number of sequences on the inference accuracy was tested. For this purpose model trees with 10, 20, 30, 40 and 50 taxa were constructed using NJ. In order to construct a model tree with N sequences, the N most divergent sequences were selected.

Table 2. Simulation results: accuracy of site-specific rate inference based on models with a different number of gamma components

Dataset	NS[a]	Mean MRAD		
		M1	M2[b]	M3[c]
3adk	10	0.891	**0.790** ($P < 0.0005$)	0.770 ($P = 0.093$)
	20	0.570	**0.490** ($P < 0.0001$)	**0.475** ($P < 0.005$)
	30	0.429	**0.365** ($P < 0.0001$)	**0.356** ($P < 0.01$)
	40	0.374	**0.316** ($P < 0.0001$)	0.311 ($P = 0.19$)
	50	0.351	**0.305** ($P < 0.0001$)	0.304 ($P = 0.25$)
3pgk	10	0.775	**0.740** ($P < 0.0001$)	**0.728** ($P < 0.01$)
	20	0.534	**0.502** ($P < 0.0001$)	**0.493** ($P < 0.0005$)
	30	0.422	**0.395** ($P < 0.0001$)	**0.388** ($P < 0.0001$)
	40	0.382	**0.359** ($P < 0.0001$)	**0.352** ($P < 0.0005$)
	50	0.349	**0.330** ($P < 0.0001$)	**0.324** ($P < 0.0005$)

[a]Number of sequences.
[b]P-value between M2 and M1.
[c]P-value between M3 and M2.
Values are shown in boldtype if the difference between M2(M3) and M1(M2) is significant ($P < 0.05$).

6.2 Simulation results

A comparison between the accuracy of site-specific rate inference as a function of the number of sequences under the M1, M2 and M3 models is shown in Table 2. For a given model, our simulations show that the accuracy increases as the number of sequences increases. This finding is expected since more data are available at each site for rate inference.

It is not clear that more complex models will result in better rate inference. When comparing the effect of the number of components on the accuracy of rate estimates, two contradicting factors may operate. When the amount of data is large, rate inference is accurate for all models, and using a rich model may not be justified. However, when the number of sequences increases, the use of models with additional free parameters is more justified (Fig. 3).

Comparing M2 and M1, the first was always significantly more accurate than the latter (for the two datasets, regardless of the dataset size). M3 was significantly better than M2 for the 3pgk dataset regardless of dataset size. The difference between M3 and M2 for the 3adk dataset is more complex. Although M3 is more accurate than M2 with every number of sequences tested, the differences between the models is significant only for the intermediate-sized data (20 and 30 sequences). With 40 and 50 sequences, the difference between M3 and M2 is not significant, probably because of the substantial data size that renders inference with the M2 model sufficiently accurate. An additional increase in data size is expected to reduce the difference between the two models even further. In the case of 10 sequences, there are probably not enough data to accurately estimate the parameters of the M3 model, which can explain why the rates estimated by M3 were not significantly more

accurate than those of M2. Indeed, the log-likelihood difference between M3 and M2 for the case of 10 sequences was not statistically significant (LRT test) in all 20 simulation runs (data not shown).

7 DISCUSSION

Evolutionary models may provide misleading results if their underlying assumptions are violated (Whelan *et al*., 2001). With the dramatic increase in sequence-data availability, we are now in a position to suggest models that better mimic known biological processes and patterns. By using realistic models, we may be able to remove possible sources of error caused by oversimplified assumptions. In this study we explored a model that better describes the among-site rate variation characteristic of molecular sequence evolution. We showed that by using a mixture of a few gamma distributions the model fit the data significantly better than the commonly used single-gamma model.

The G + I model can be regarded as the first, albeit restricted, gamma mixture approach. Although the G + I model is intuitively very appealing, the estimates of the model parameters are highly sensitive to taxon sampling (Yang, 1996; Sullivan *et al*., 1999). In addition, the high correlation between the proportion of the invariable sites and the gamma shape parameter indicates model inadequacy (Sullivan *et al*., 1999). Our results here indicated that, in term of AIC_c scores, none of the datasets examined could be best explained by the G + I model, whereas the M1, M2 and M3 models best fitted 1, 7 and 2 datasets, respectively.

Recently, a few studies have suggested alternatives to the single-gamma model. Yang *et al*. (2000) have studied codon substitutions models with a range of different distributions for modeling the non-synonymous/synonymous rate ratio among sites, including a two-gamma distribution (model M6). Kosakovsky Pond and Frost (2005) have suggested a hierarchical model, in which the baseline gamma distribution is discretized into several categories using a beta distribution. Using AIC scores it was shown that this model better fits several nucleotide coding datasets. Susko *et al*. (2003) investigated a non-parametric model that does not involve a prior on the rate distribution. Instead, the rate distribution is estimated using a large number of free parameters. Although these two models may capture deviations from the gamma distribution, the mixture model approach suggested in our study adds an additional flexibility since the exact number of components needed to capture the complexity of the rate distribution is adjustable and can be statistically inferred. As an alternative to using a mixture of gammas, a mixture of other positive distributions, such as the log-normal, can be applied using a similar mechanism described in this study. Clearly, more analyses are needed in order to study the merits of these different models under a range of evolutionary scenarios.

Phylogenetic reconstructions based on large datasets are becoming increasingly popular. For example, Murphy *et al*. (2001) investigated placental phylogeny using 16.4 kb molecular data for 44 taxa. Reyes *et al*. (2004) used the complete mitochondrial genomes of 77 species to reconstruct a mammalian phylogeny. Large molecular datasets have the potential to resolve longstanding controversies in systematics. In such analyses, the data may support accurate parameter estimates of complex models. Using a mixture of gamma approach is a natural extension of the single-gamma model, allowing a better fit to the distribution of evolutionary rates yet without introducing too many free parameters. Our results showed that the M2 model was superior to M1 for all data analyzed (Table 1). We also showed that the better fit of M2 increases as larger datasets are analyzed (Fig. 3). This suggests that the gamma mixture model will be particularly valuable for such large datasets analyses.

The EM algorithm used in the optimization process only guarantees to find a local maximum, rather than a global one. Unfortunately, local optima are a practical difficulty, especially when the number of parameters increases. Initiating the EM algorithm from several random points of the parameter space is needed to avoid local optima. Our practical experience suggests that with M1, the EM algorithm constantly reaches the global maxima (as a few dispersed starting points reach indistinguishable results). However, as the number of components increases, the EM algorithm often terminates at local optima. This implies that the observed superiority of models with more gamma components is underestimated.

Discretization of the continuous gamma distribution can be better achieved using the Laguerre quadrature method compared with Yang's (1994) quantile approximation (Felsenstein, 2001; Kosakovsky Pond and Frost, 2005). The Laguerre quadrature method was used here not only to compute the likelihood of a tree, but also for the various numerical integrations that are a part of the EM algorithm. Using the Laguerre quadrature we obtained faster and more consistent results than the more commonly used quantile approach (data not shown).

Knowledge of relative evolutionary rates is crucial not only to define more appropriate evolutionary models, but also because it serves as a means to evaluate the importance of a site in maintaining the structure or function of a protein. Furthermore, covarion models can be used to detect sites that exhibit different evolutionary rates in different branches of the phylogenetic tree. Such rate shifts may indicate change in the selection intensity at specific sites during evolution (Knudsen and Miyamoto, 2001; Gaucher *et al*., 2002; Pupko and Galtier, 2002; Susko *et al*., 2002). We have recently shown that an empirical Bayesian method produces significantly more accurate rate estimations than an ML method, indicating that the gamma prior that is integrated into the Bayesian computation significantly improves performance (Mayrose *et al*., 2004). Here we further showed that the incorporation of a gamma mixture prior leads to even better rate inferences.

ACKNOWLEDGEMENTS

We thank D. Burstein, A. Doron-Faigenboim, M. Ninio, E. Privamn, N. Rubinstein and A. Stern for their insightful comments. This work is supported in part by ISF grant number 1208/04 to N.F. and T.P. T.P. was also supported by a grant in Complexity Science from the Yeshaia Horvitz Association.

REFERENCES

Abramowitz,M. and Stegun,I.A. (1972) *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tablesprinting*. Dover, New York.

Anisimova,M. *et al*. (2001) Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol. Biol. Evol.*, **18**, 1585–1592.

Buntine,W. (1994) Operations for learning with graphical models. *J. Artif. Intel. Res.* **2**, 159–225.

Burnham,K.P. and Anderson,D.R. (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer-Verlag, New York, NY.

Dempster,A.P. *et al*. (1977) Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B.*, **39**, 1–38.

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

Felsenstein,J. (2001) Taking variation of evolutionary rates between sites into account in inferring phylogenies. *J. Mol. Evol.*, **53**, 447–455.

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

Gaucher,E.A. *et al*. (2002) Predicting functional divergence in protein evolution by site-specific rate shifts. *Trends Biochem. Sci.*, **27**, 315–321.

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

Jones,D.T. *et al*. (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.*, **8**, 275–282.

Knudsen,B. and Miyamoto,M.M. (2001) A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. *Proc. Natl Acad. Sci. USA*, **98**, 14512–14517.

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

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.

Murphy,W.J. *et al*. (2001) Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science*, **294**, 2348–2351.

Press,W.H., Teukolsky,S.A., Vetterling,W.T. and Flannery,B.P. (2002) *Numerical Recipes in C++: The Art of Scientific Computing*. Cambridge University Press, Cambridge.

Pupko,T. and Galtier.N. (2002) A covarion-based method for detecting molecular adaptation: application to the evolution of primate mitochondrial genomes. *Proc. R. Soc. Lond. B. Biol. Sci.*, **269**, 1313–1316.

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**, S71–S77.

Reyes,A. *et al*. (2004) Congruent mammalian trees from mitochondrial and nuclear genes using Bayesian methods. *Mol. Biol. Evol.*, **21**, 397–403.

Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406-425.

Sander,C. and Schneider,R. (1993) The HSSP data base of protein structure-sequence alignments. *Nucleic Acids Res.*, **21**, 3105–3109.

Sokal,R.R. and Rohlf,F.J. (1981) *Biometry: The Principles and Practice of Statistics in Biological Research*. W.H. Freeman and Company, New York, NY, pp. 447–450.

Sullivan,J. *et al*. (1999) The effect of taxon sampling on estimating rate heterogeneity parameters of maximum-likelihood models. *Mol. Biol. Evol.*, **16**, 1347–1356.

Susko,E. *et al*. (2002) Testing for differences in rates-across-sites distributions in phylogenetic subtrees. *Mol. Biol. Evol.*, **19**, 1514–1523.

Susko,E. *et al*. (2003) Estimation of rates-across-sites distributions in phylogenetic substitution models. *Syst. Biol.*, **52**, 594–603.

Sussman,J.L. *et al*. (1998) Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta. Crystallogr. D. Biol. Crystallogr.*, **54**, 1078–1084.

Swofford,D.L., Olsen,G.J., Waddell,P.J. and Hillis,D.M. (1996) Phylogenetic inference. In: Hillis,D.M., Moritz,C. and Mable,B.K. (eds), *Molecular Systematics*. Sinauer Associates, Sunderland, MA.

Whelan,S. *et al*. (2001) Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends Genet.*, **17**, 262–272.

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. *et al*. (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, **155**, 431–449.

APPENDIX: A LAGUERRE QUADRATURE METHOD FOR APPROXIMATING THE GAMMA DISTRIBUTION

Felsenstein (2001) suggested that the Laguerre quadrature method for the discretization of the gamma distribution can better approximate the continuous distribution compared with Yang's (1994) quantile approximation. We used the Laguerre method not only to compute the likelihood of a tree, but also for the various numerical integrations that are a part of the EM algorithm. For the EM we approximate a gamma distribution with both α and β parameters. We, thus, give a detailed explanation of this approximation.

When approximating a continuous distribution by a discrete one [Equation (5)], discrete values (rates) $\tilde{r}_i$ and probabilities $w(\tilde{r}_i)$ must be chosen so as to approximate the integral most accurately. The idea of Gaussian quadrature is to give ourselves the freedom to choose not only the location of the abscissas at which the function is to be evaluated, as in Yang's (1994) discrete approximation, but also the weighting coefficients. Specifically, we want to approximate the integral $\int_a^b W(x)f(x)dx$, where $W(x)$ is called the weighting function. Given an integer N we can find weights w_j and abscissas x_j such that the approximation

$$\int_a^b W(x)f(x)dx \cong \sum_{j=1}^{N} w_i f(x_i) \qquad \text{(A1)}$$

is exact, if $f(x)$ is a polynomial of degree $2N-1$ or less. The x_j are the roots of a set of orthogonal polynomials that depend on the weighting function $W(x)$. Once the abscissas are determined, the weights w_j can be found (w_j are constructed such that their sum equals 1). The problem is usually how to construct the associated set of orthogonal polynomials. Different numerical quadrature methods depend on which weighting function is used.

Fortunately, the gamma density function can be converted into a classical weighting function $W(m) = \mathrm{e}^{-m}m^{\alpha}$. The abscissas and weights of this weighting function can be easily calculated using the generalized Laguerre quadrature method (Press *et al*., 2002). In what follows we detail how to convert between these two functions.

We want to approximate the integrals, such as those presented in Equations (1), (17) and (18)

$$\int_0^{\infty} g(r:\alpha,\beta)f(r)dr = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\int_0^{\infty} \mathrm{e}^{-\beta r}r^{\alpha-1}f(r)dr. \qquad \text{(A2)}$$

By setting $m = \beta r$ we get

$$\frac{1}{\Gamma(\alpha)}\int_0^{\infty} \mathrm{e}^{-m}m^{\alpha-1}f\left(\frac{m}{\beta}\right)dm. \qquad \text{(A3)}$$

By setting $\alpha' = \alpha - 1$, we obtain the desired weighting function. To conclude, given N, the approximation is

$$\int_0^{\infty} g(r:\alpha,\beta)f(r)dr \cong \sum_{j=1}^{N} f(\tilde{r}_j)w(\tilde{r}_j), \qquad \text{(A4)}$$

where $w(\tilde{r}_j) = \frac{w_j}{\Gamma(\alpha)}$, $\tilde{r}_j = \frac{m_j}{\beta}$. The terms m_j and w_j are the roots and weights of the Gauss–Laguerre quadrature formula with $\alpha' = \alpha - 1$.