Modeling Protein Evolution with Several Amino Acid Replacement Matrices Depending on Site Rates

Si Quang LE[1,2], Cuong Cao DANG[3], and Olivier GASCUEL[1*]

1 : Méthodes et Algorithmes pour la Bioinformatique

LIRMM, CNRS - Université Montpellier II,

161 rue Ada, 34392 – Montpellier Cedex 5 – France

Tel. 33 (0) 4 67 41 85 47 – Fax. 33 (0) 4 67 41 85 00

URL: http://www.lirmm.fr/mab

2: Wellcome Trust Sanger Institute, Genome Campus, Hinxton, UK.

3: University of Engineering and Technology, Vietnam National University, Hanoi.

Emails: le@lirmm.fr, gascuel@lirmm.fr

* Corresponding author

Abstract

Most protein substitution models use a single amino acid replacement matrix summarizing the biochemical properties of amino acids. However, site evolution is highly heterogeneous and depends on many factors that influence the substitution patterns. In this paper we investigate the use of different substitution matrices for different site evolutionary rates. Indeed, the variability of evolutionary rates corresponds to one of the most apparent heterogeneity factors among sites, and there is no reason to assume that the substitution patterns remain identical regardless of the evolutionary rate. We first introduce LG4M, which is composed of four matrices, each corresponding to one discrete gamma rate category (out of four). These matrices differ in their amino acid equilibrium distributions and in their exchangeabilities, contrary to the standard gamma model where only the global rate differs from one category to another. Next, we present LG4X, which also uses four different matrices, but leaves aside the gamma distribution and follows a distribution-free scheme for the site rates. All these matrices are estimated from a very large alignment database, and our two models are tested using a large sample of independent alignments. Detailed analysis of resulting matrices and models shows the complexity of amino acid substitutions and the advantage of flexible models such as LG4M and LG4X. Both significantly outperform single-matrix models, providing gains of dozens to hundreds of log-likelihood units for most data sets. LG4X obtains substantial gains compared to LG4M, thanks to its distribution-free scheme for site rates. Since LG4M and LG4X display such advantages but require the same memory space and have comparable running times to standard models, we believe that LG4M and LG4X are relevant alternatives to single replacement matrices. Our models, data and software are available from http://www.lirmm.fr/~le/LG4X/.

Keywords: amino-acid substitutions, replacement matrices, gamma and distribution-free rate models, maximum likelihood estimations, phylogenetic inference

Introduction

Amino acid replacement matrices ─20x20 matrices containing estimates of the instantaneous substitution rates of any amino acid by another─ are essential in most methods to infer protein phylogenies. These matrices are expected to capture the biological and physico-chemical properties of amino acids. They are used in distance-based methods to estimate the evolutionary distance ─the expected number of substitutions per site─ between sequence pairs. In maximum likelihood and Bayesian methods, they are used to compute substitution probabilities along tree branches, and hence the likelihood of the data (see textbooks, e.g., Felsenstein 2003; Yang 2006).

The standard approach to infer protein phylogenies is based on the use of a single replacement matrix. Several general matrices estimated from very large sets of taxa and alignments have been proposed since the pioneering work of Dayhoff, Eyck and Park (1972), notably JTT (Jones, Taylor and Thornton 1992), WAG (Whelan and Goldman 2001) and LG (Le and Gascuel 2008). Some studies showed that specific matrices should be used for certain analyses, for example with membrane (Jones, Taylor and Thornton 1994) or mitochondrial (Yang, Nielsen and Hasegawa 1998) proteins, but general matrices are usually robust and tend to perform well in many cases (Keane et al. 2006). However, site evolution is highly heterogeneous and depends on many factors such as genetic code, solvent accessibility, secondary and tertiary structure, and protein functions. Most notably, some sites are subject to strong evolutionary pressure and evolve slowly due to their role in the structure or functions of the protein, while others are much less constrained and accumulate substitutions rapidly. In the standard approach, this variability is modeled by discrete gamma rate categories, which are used to modulate the (unique) replacement matrix being selected, depending on the site rates (Yang 1993). As site rates are unknown, all rates are envisaged for every site and accounted for thanks to a mixture approach (see textbooks, e.g., Gascuel and Guindon 2007).

However, many works revealed that depending on site specificities, not only do the global rates vary, but also the substitution patterns. Notably, buried sites (typically slow) and

exposed sites (typically fast) obey very different matrix models (Koshi and Goldstein 1995; Lio et al. 1998; Goldman, Thorne and Jones 1998; Holmes and Rubin 2002; Le, Lartillot and Gascuel 2008; Le and Gascuel 2010). To a lesser extent, it was also shown that substitution processes vary among secondary structures (Koshi and Goldstein 1995; Thorne, Goldman and Jones 1996; Le, Lartillot and Gascuel 2008; Le and Gascuel 2010). All of these works (and others) thus explored site-dependent models using several matrices or profiles.

In the profile approach (Koshi and Goldstein 1998; Lartillot and Philippe 2004; Le, Gascuel and Lartillot 2008), sets of elementary models defined by their amino acid equilibrium frequencies are used; these models rely on simple multinomial processes over the 20 amino acids ─analogous to the (Felsenstein 1981) model of DNA substitution─ and do not use replacement matrices (or only highly simplified ones). In the multi-matrix approach (Koshi and Goldstein 1995; Thorne, Goldman and Jones 1996; Goldman, Thorne and Jones 1998; Le, Lartillot and Gascuel 2008), different matrices are used for different site categories. The model introduced by Wang et al. (2008) is a compromise between these two approaches as it uses several (full range) matrices that only differ in their amino acid equilibrium distributions (see also Lartillot and Philippe 2004, for a similar model). In all cases, the set of profiles or matrices is combined thanks to a mixture approach or a HMM (Felsenstein and Churchill 1996; Thorne, Goldman and Jones 1996). In recent studies (Le, Gascuel and Lartillot 2008; Le, Lartillot and Gascuel 2008; Wang et al. 2008) this first-level mixture is combined with a second-level mixture corresponding to the standard gamma rate categories. This combination was shown to be quite accurate, but is computationally heavy as both the computing time and the memory consumption are roughly proportional (see, e.g., Bryant, Galtier and Poursat 2005) to the number of site categories (e.g., 12 with four gamma categories and three biochemical categories).

In this paper we investigate simpler models, where sites are categorized depending on their evolutionary rate, and different replacement matrices are used for each site category. Indeed, the variability of evolutionary rates corresponds to the one of most apparent heterogeneity factor among sites, and there is no reason to suppose (as in the standard

approach) that the substitution pattern remains identical regardless the evolutionary rate. For example, we expect slow sites to be mostly hydrophobic (and fast sites to be hydrophilic), which implies that the amino-acid equilibrium frequencies should vary depending on the site rate. Investigated models thus focus on an essential site heterogeneity factor. They refine the standard gamma model by using several different replacement matrices, instead of only one modulated by a global rate. However, these models are less complex than two-level mixtures as they use a single mixture level enabling fair computing times and low memory consumption.

We first verify the use of different matrices for different evolutionary rates that follow a discrete gamma distribution. To this end, we estimate a 4-matrix model (LG4M), where each matrix corresponds to one standard gamma rate category (Γ4; Yang 1993). Experimental results show that LG4M outperforms single-matrix models (JTT+Γ4, WAG+Γ4 and LG+Γ4) in terms of tree likelihood, and often infers different tree topologies.

We then examine the limitation of constraining rates to a gamma distribution by testing a model of four matrices where rates and weights of the four matrices are freely estimated. To this end, we estimate another 4-matrix model (LG4X), where rates and weights are left out of the gamma distribution assumption. Experimental results show that LG4X is significantly better than LG4M, and comparable to the two-level mixture models from (Le, Lartillot and Gascuel 2008; Le and Gascuel 2010), and at the same time much simpler.

These results, combined with low computing times and memory consumption, suggest that LG4M and LG4X are relevant alternatives to standard single-matrix models in inferring phylogenetic trees from protein sequences. In the following, we first describe our data, then our models and their estimation procedures, and lastly provide comparisons with independent testing alignments.

Data Sets

To estimate LG4M and LG4X, we used alignments extracted from HSSP (Schneider et al. 1997). This database comprises ~50,000 alignments of protein families, each containing an

average of ~550 members. Each alignment is obtained by aligning a protein with known three-dimensional structure in the Protein Data Bank (PDB; Berman et al. 2000) to all its probable sequence homologs in UNIPROT. The protein with known structure is called the 'test protein' of the alignment. HSSP alignments contain a huge number of gaps due to absent or unsequenced domains for some proteins. Consequently, we cleaned each alignment by selecting sequences that were well-aligned, sufficiently different one from the other, and had 40-99% identities with the test protein. Gapped regions among selected sequences were eliminated using GBLOCKS (Castresana 2000) with default options, and we removed alignments with less than 10 selected sequences or 100 remaining sites. We also left out membrane proteins (based on their presence in the Membrane Protein Data Bank, Raman, Cherezov and Caffrey 2006) since their amino acid replacement pattern is highly different to that of globular proteins (Jones, Taylor and Thornton 1994). Moreover, HSSP is highly redundant because a protein sequence may appear in more than one alignment depending on its homologs with known structure in PDB. Thus, we retained only independent alignments that do not share any sequence. To this end, we used a heuristic algorithm to find a large number of independent alignments containing a large number of sites with few gaps (Le, Lartillot and Gascuel 2008). This selection procedure resulted in 1,771 non-redundant alignments, with an average of ~56 sequences and ~254 sites per alignment, a total of ~27 million amino acids and less than 0.1% gaps. We randomly picked 1,471 alignments to estimate LG4M and LG4X, and used the remaining 300 for model comparison. These alignments were the same as those used to estimate and test our two-level mixtures of profiles and matrices, and our structure-informed models (Le, Gascuel and Lartillot 2008; Le, Lartillot and Gascuel 2008; Le and Gascuel 2010). Additional details on the selection procedure are provided in these references, and the training and test alignments are available from www.lirmm.fr/~le/LG4X.

To assess the performance of our models, we used the 300 HSSP test alignments and another set of independent alignments extracted from TreeBase (Sanderson et al. 1994). This database contains alignments that were produced especially for phylogenetic analyses, and

thus provide a good benchmark for comparing models meant for phylogenetic reconstruction. Moreover, the use of test alignments from a different database should avoid possible biases induced by some feature specific to our HSSP training alignments. We took all (113) most recently updated TreeBase globular protein alignments and then removed those including too many gaps (>45%) or showing a too high level of sequence divergence (average number of amino acids per site >8, presence in the ML tree of one or several branches with length >2.0, or average ML tree branch length >0.50). We retained 84 alignments, with size ranging from small, single protein alignments (e.g., 7 taxa and 232 sites), to very large concatenated protein alignments (e.g., 62 taxa and 11,544 sites). These TreeBase test alignments are also available from www.lirmm.fr/~le/LG4X.

Models

All amino acid substitution matrices discussed here comply with the general time-reversible (GTR) model (see textbooks, e.g., Bryant, Galtier and Poursat 2005; Yang 2006). Such a matrix is denoted $Q=\left(q_{xy}\right)$, where q_{xy} is the substitution rate from amino acid x to amino acid $y\ (\neq x)$; diagonal terms are set such that the row sums are all zero, that is, $q_{xx}=-\sum_{y\neq x}q_{xy}$. Thanks to time reversibility, Q can be decomposed into the symmetric exchangeability matrix $R=\left(r_{x\leftrightarrow y}\right)$ and the amino acid equilibrium distribution $\pi=\left(\pi_x\right)$, using $q_{xy}=\pi_y r_{x\leftrightarrow y}\ (x\neq y)$. The amino acid distribution (π) may be estimated from the training alignments and is then called the model equilibrium distribution, or from the data analyzed (+F option). With single-matrix models, Q and R are normalized such that one time unit corresponds to one substitution per site, that is, $\rho=-\sum_x Q_{xx}\pi_x=1.0$, where ρ is the global rate of Q. This constraint is released with some multi-matrix models (e.g., Le, Lartillot and Gascuel 2008), where some site categories and matrices are fast with a high global rate and some others are slow with a low ρ value. Here we use normalized matrices only, but modulate their global rate using external parameters with values fitted on the analyzed alignment (see below). A matrix Q then contains 208 free parameters (190 in R + 19 in π - 1 normalization constraint).

The likelihood of the data (denoted D) for a given tree T (including branch lengths) and replacement matrix Q is

$$L(T,Q;D) = \prod_i L(T,Q;D_i), \tag{1}$$

where the product runs over all the sites (independence assumption), and $L(T,Q;D_i)$ is the likelihood of the data at site i (D_i) given T and Q. $L(T,Q;D_i)$ is computed efficiently thanks to the the pruning algorithm (Felsenstein 1981).

Yang (1993) introduced a mixture model based on a single replacement matrix but variable rates across sites following a discrete gamma distribution with K equally weighted rate categories. With $K = 4$ the data likelihood is given by

$$L(T,Q,\alpha;D) = \prod_i \left(\frac{1}{4} \sum_{k=1}^{4} L(T,\Gamma(\alpha,k)Q;D_i) \right), \tag{2}$$

where $\Gamma(\alpha,k)$ is the k^{th} rate of a discrete gamma distribution with parameter α. The weights (or contributions) of rate categories are all equal to $1/K$. Both T and α are estimated by maximizing likelihood (2). Variants of this model include non equally weighted gamma rate categories (e.g., Susko et al. 2003; Mayrose, Friedman and Pupko 2005), an approach that could be further investigated to improve models presented here.

Multi-matrix models were proposed by several authors (e.g., Koshi and Goldstein 1995; Thorne, Goldman and Jones 1996; Goldman, Thorne and Jones 1998) to account for the secondary structure and solvent accessibility. With such models, the data likelihood in a mixture context is expressed as

$$L(T,\mathrm{Q} = \{Q_1,..,Q_M\}, \mathrm{w} = \{w_1,...,w_M\}; D) = \prod_i \left(\sum_{m=1}^{M} w_m L(T,Q_m;D_i) \right), \tag{3}$$

where M is the number of matrices, and w_m the weight of matrix Q_m, with constraint $\sum_{m=1}^{M} w_m = 1$.

Recent works (e.g., Le, Lartillot and Gascuel 2008; Wang et al. 2008) combined Yang's model (2) with the above (3) multiple-matrix model

$$L(T,\mathrm{Q}=\{Q_1,..,Q_M\},\mathrm{w}=\{w_1,...,w_M\},\alpha;D)=\prod_i\left(\sum_{m=1}^{M}\frac{w_m}{K}\sum_{k=1}^{K}L(T,\Gamma(\alpha,k)Q_m;D_i)\right), \quad (4)$$

where constraint $\sum_{m=1}^{M} w_m = 1$ still holds. Equation (4) expresses two levels of mixture, one for gamma distributed rate categories, and one for multiple substitution matrices. In this framework we introduced several supervised and unsupervised models, for example (supervised) EX2 with two matrices for buried and exposed sites, and (unsupervised or 'blind') UL3 based on three matrices that were estimated without *a priori* knowledge on site categorization (Le, Lartillot and Gascuel 2008). The same framework was used by Wang et al. (2008) in a 5-matrix model, where all matrices were based on the same JTT or WAG exchangeability matrix (R), but used different amino acid equilibrium distributions (π).

Although the above models (EX2, UL3, etc.) perform well and provide high likelihood values, they are computationally expensive in terms of both computing time and memory consumption. This is mainly due to their high number of site categories, for example twelve with UL3 and four gamma categories. In this paper, we explore simplifications of Equation (4). In our LG4M model, we assume four equally weighted gamma rate categories, and use four matrices, one for each rate category. Let $\mathrm{Q}=\{Q_1,Q_2,Q_3,Q_4\}$ be the set of these four matrices, where Q_1 stands for the matrix corresponding to slowest category and Q_4 that of the fastest one. Data likelihood is expressed as

$$L(T,\mathrm{Q},\alpha;D)=\prod_i\left(\frac{1}{4}\sum_{k=1}^{4}L(T,\Gamma(\alpha,k)Q_k;D_i)\right). \quad (5)$$

Mathematically speaking, this model (5) is a compromise between Yang's model (2) and two-level mixture models (4). Instead of sharing the same matrix as in Yang's model, each rate has its own matrix, and each matrix is applied only to one rate category instead of being applied to all rates as in two-level mixture models. This model (5) is thus more general than Yang's model, but keeps the same free parameters to be estimated from the data (i.e. α and T)

as in Yang's model. From a biological standpoint, the simplification from two-level mixture (4) to model (5) means that the main heterogeneity factor among sites is their evolutionary rate, an assumption that will be tested using independent alignments in the Performance Comparison section.

Model LG4M in Equation (5) constrains the site rates using a discrete gamma distribution. In our LG4X model, we generalize LG4M by removing this constraint

$$L(T,\mathrm{Q},\rho=\{\rho_1,\rho_2,\rho_3,\rho_4\},\mathrm{w}=\{w_1,w_2,w_3,w_4\};D)=\prod_i\left(\sum_{k=1}^{4}w_k L(T,\rho_k Q_k;D_i)\right), \quad (6)$$

where w_k and ρ_k are the weight and rate of matrix Q_k such that $\sum_{k=1}^{4}w_k=1$ and $\sum_{k=1}^{4}w_k\rho_k=1$. The latter normalization constraint is needed to get 1.0 substitution per site within one time unit, just as in standard single-matrix models (this normalization is implicit in LG4M). This model thus involves three free parameters among weights w_k, plus three free parameters among rates ρ_k, which are estimated by maximizing likelihood (6) on the data set analyzed.

Model Estimation

We have a set of N protein alignments denoted $\mathrm{D}=\{D^1,...,D^N\}$, where D^a is an alignment. We aim to estimate a 4-matrix model $\mathrm{Q}^*=\left(Q_1^*,Q_2^*,Q_3^*,Q_4^*\right)$ that maximizes the likelihood of D

$$\mathrm{Q}^*=\underset{\mathrm{Q}=(Q_1,Q_2,Q_3,Q_4),\mathrm{T},\mathrm{P},\mathrm{W}}{\arg\max}\left\{\prod_{a=1}^{N}L\left(T^a,\mathrm{Q},\rho^a,\mathrm{w}^a;D^a\right)\right\}, \quad (7)$$

where $\mathrm{T}=(T^1,...,T^N)$, $\mathrm{P}=(\rho^1,...,\rho^N)$ and $\mathrm{W}=\left(\mathrm{w}^1,...,\mathrm{w}^N\right)$ are the trees, rates and weights of the N alignments, respectively; $L\left(T^a,\mathrm{Q},\rho^a,\mathrm{w}^a;D^a\right)$ is the likelihood of D^a given model Q, tree T^a, rates $\rho^a=(\rho_1^a,...,\rho_4^a)$, and weights $\mathrm{w}^a=(w_1^a,...,w_4^a)$. Thus, to estimate $\mathrm{Q}^*=\left(Q_1^*,Q_2^*,Q_3^*,Q_4^*\right)$ we also need to estimate T^*, P^* and W^* which optimize likelihood (7). We are interested in two 4-matrix models: LG4M where $L\left(T^a,\mathrm{Q},\rho^a,\mathrm{w}^a;D^a\right)$ is calculated using Equation (5), and LG4X that is based on Equation (6). For each alignment D^a, ρ^a and

w^a of LG4M follow a discrete gamma distribution with four equally-weighted rate categories, while in LG4X, these parameters are freely estimated such that $\sum_1^4 w_k^a = 1$ and $\sum_1^4 w_k^a \rho_k^a = 1$, without any additional constraint.

Following (Whelan and Goldman 2001; Le and Gascuel 2008; Le, Lartillot and Gascuel 2008), we estimate $\mathrm{Q}^* = \left(Q_1^*, Q_2^*, Q_3^*, Q_4^*\right)$ from Equation (7) in two steps: (i) given a fixed starting value of Q, we estimate T^*, P^* and W^* by maximizing likelihood (5) (LG4M) or (6) (LG4X); (ii) we then estimate $\mathrm{Q}^* = \left(Q_1^*, Q_2^*, Q_3^*, Q_4^*\right)$ using Equation (7) with respect to T^*, P^* and W^* values obtained in step (i). These two steps are iterated one after the other until no more improvement of T^*, P^*, W^* and Q^* is found.

Since trees, rates, and weights of alignments in D are independent of one another, we optimize T^*, P^* and W^* for each alignment D^a independently

$$\forall D^a : (T^a, \rho^a, \mathrm{w}^a) = \arg\max_{T,\rho,\mathrm{w}} \left\{ L\left(T, \mathrm{Q}, \rho, \mathrm{w}; D^a\right) \right\}. \tag{8}$$

For this purpose, we use an adaptation of PhyML 3.0 (Guindon et al. 2010) that is described below. Having obtained T^*, P^* and W^*, we search for Q^* that maximizes the likelihood of the data given T^*, P^* and W^*

$$\mathrm{Q}^* = \left(Q_1^*, Q_2^*, Q_3^*, Q_4^*\right) = \arg\max_{\mathrm{Q}=(Q_1,Q_2,Q_3,Q_4)} \left\{ \prod_{a=1}^{N} L\left(T^a, \mathrm{Q}, \rho^a, \mathrm{w}^a; D^a\right) \right\}. \tag{9}$$

It is impractical to optimize $\left(Q_1^*, Q_2^*, Q_3^*, Q_4^*\right)$ directly from Equation (9) due to the huge number (4 x 208) of free parameters in Q. Consequently, we use the approximate learning method proposed in (Le and Gascuel 2008; Le, Lartillot and Gascuel 2008) where $\mathrm{Q}^* = \left(Q_1^*, Q_2^*, Q_3^*, Q_4^*\right)$ is handled by simplifying the site likelihood in Equation (9) using the site rate category with maximum posterior probability (MAP) only, instead of summing overall rate categories, that is,

$$\mathrm{Q}^* = \left(Q_1^*, Q_2^*, Q_3^*, Q_4^*\right) = \arg\max_{\mathrm{Q}=(Q_1,Q_2,Q_3,Q_4)} \left\{ \prod_a \prod_i L\left(T^a, \rho_{c_i}^a Q_{c_i}; D_i^a\right) \right\}, \tag{10}$$

where D_i^a is the *i*th site of alignment D^a, c_i is the MAP rate category (computed during tree estimation) for site D_i^a, and ρ_{c_i} is the rate of c_i corresponding to Q_{c_i} substitution matrix. Equation (10) can then be rewritten as

$$\forall k = 1..4,\ Q_k^* = \arg\max_{Q_k} \left\{ \prod_a \prod_{i:c_i=k} L\left(T^a, \rho_k^a Q_k; D_i^a\right) \right\}. \quad (11)$$

In other words, every Q_k is estimated independently. To achieve these estimations, we used XRate (Holmes and Rubin 2002; Klosterman et al. 2006) with the same search options as in (Le and Gascuel 2008; Le, Lartillot and Gascuel 2008). Notably, we used the forgiven option (with 3 jumps) to escape from local optima. XRate is able to deal with mixtures, instead of using our simplifying MAP-based approach (11). However, we observed that using MAP in this estimation context is much faster, less affected by local optima, and tends to provide better results (Le and Gascuel 2008). This is why here we adopted the same strategy, which is close to Viterbi's approximation that proved to be both efficient and accurate when estimating HMMs (Durbin et al. 1998).

To perform these computations and use our new models to infer trees, we adapted PhyML 3.0 (Guindon et al. 2010) to LG4X and LG4M. This dedicated version is called PhyML-4X in the following. The adaptation of PhyML to LG4M is just a trigger so that the program selects the correct matrix for each rate category. The other parts (for example to optimize α or to search tree topologies) are kept the same as in standard PhyML. In the case of LG4X, we reused the optimization module from (Le, Lartillot and Gascuel 2008) to optimize weights $(w_1,...,w_4)$ and rates $(\rho_1,...,\rho_4)$, alternating mono-dimensional Brent optimization of every variable until global convergence. To account for constraint $\sum w_m = 1$, we use the variable change $w_i = e^{v_i} / \sum e^{v_m}$ and then optimize the v_is using Brent. The second constraint $\sum w_m \rho_m = 1$ is fulfilled by rescaling rates and branch lengths before returning the final tree. To accelerate the calculations, the tree and the starting rate and weight (= 0.25) values are estimated with LG4M, which involves a single (α) parameter to be optimized instead of six.

Figure 1 summarizes the whole estimation procedure. Both LG4M and LG4X are initialized starting from the LG matrix. LG4M uses a supervised approach where each matrix is associated with the same gamma rate category throughout the optimization procedure; for example, the 'Fast' matrix is systematically associated with the highest rate, among four gamma-distributed rates. LG4X is estimated in a semi-supervised way. During the first step, sites are categorized based on the rate (associated to LG) providing the highest likelihood value. During subsequent steps, sites are categorized based on the (rate, matrix) pair with highest likelihood. In most cases, the 'Fast' matrix is associated with the highest rate, and the same holds with other matrices. However, since rates (and weights) are estimated independently for each alignment without any *a priori* constraint, it may occur for some datasets that the 'Fast' matrix is actually associated with a slow rate and *vice versa* (see below, Table 1 and Supplementary Material). Thus, this semi-supervised procedure provides a measure of the importance of the site rate factor. If the initial rate-based site categorization and matrix interpretation disappeared during the subsequent training steps, this would mean that site rate is not a heterogeneity factor of first importance, and that other more important factors exist. If (as is the case), the initial rate-based site categorization and matrix interpretation is (mostly) preserved all along training steps, this implies that the rate factor is of first importance, as we assumed in this study.

As all Expectation-Maximization (EM) approaches, XRate is sensitive to starting parameter values. For computing-time reasons (LG4X required nearly a week to be estimated, and much more to be tested and compared to other models), we did not try alternative starting matrices and training strategies. However, based on our previous experiments with LG where XRate performed remarkably well (Le and Gascuel 2008), we are confident that LG4M (estimated in a supervised manner) should be relatively stable and insensitive to the starting matrix used to initiate the training procedure. On the other hand, we also observed (Le, Lartillot and Gascuel 2008) that semi-supervised training of mixture models is more sensitive to the choice of the starting point. This suggests that LG4X could likely be improved using other starting points or training strategies.

LG4M and LG4X Matrices and Models

LG4M and LG4X matrices (estimated as described above) are available from Supplementary Material (www.lirmm.fr/~le/LG4X) along with additional information and statistics. Here we discuss the main features of these matrices and models that make them better than single-matrix models, especially LG4X thanks to its rate distribution-free scheme. Table 1 provides summary statistics and Figure 2 shows some illustrative matrices. It can be seen that:

LG4M and LG4X matrices clearly depart from LG. The correlation of the log-entries with those of LG (LogCor/LG, Tab. 1) is below 0.9 in most cases. LG4M 'Medium' is a noticeable exception (LogCor/LG = 0.957), which is somewhat expected as this matrix is used for intermediate sites with evolutionary rates close to 1. For each matrix, Table 1 provides its global hydropathy (Hydro), computed as the average hydropathy index (Kyte and Doolittle 1982) of the 20 amino acids with weights equal to their equilibrium frequencies in the given matrix. This index also points to a clear difference between the new matrices and LG (Hydro = -0.253). Most of the matrices (e.g., LG4X 'Fast' = -1.815 or LG4M 'Slow' = 1.249) are clearly hydrophilic or clearly hydrophobic, with hydropathy values close to that of 'Exposed' (-1.993) or 'Buried' (1.715) matrices from our EX3 model (Le, Lartillot and Gascuel 2008). These results and measures support our working hypothesis that the substitution patterns differ depending on the site rates. Modulating a unique replacement matrix (e.g., LG) using gamma distributed rates appears to be an oversimplification.

'Very Slow' matrices show a remarkable pattern, especially that of LG4X (Fig. 2) which is mostly used to express high replacement rates between amino acid pairs that are biochemically very similar, for example: R and K (positively charged), D and E (negatively charged), and F and Y (aromatic). These three pairs are very close in the genetic code, requiring only one nucleotide change to mutate amino acid into the other. Interestingly, some of these pairs are highly hydrophilic (e.g., R and K), which contradicts the first intuition that very slow sites should be all buried and hydrophobic. However, to avoid misinterpretation, it has to be noted that rates displayed in Fig. 2 are relative rates; all matrices are normalized and

do not incorporate the fact that very slow sites globally evolve slower (~6 times in average; Tab. 1) than fast sites; for example, the absolute rate (accounting for this global factor of ~6) between R and K is nearly symmetrical and almost the same in the 'Very Slow' and 'Fast' matrices of LG4X. In other words, R-K replacements are fast in all rate categories, including the slowest one. The 'Very Slow' LG4M matrix is less contrasted than the LG4X matrix and deals with other amino acid groups, also very close biochemically and in the genetic code, for example: I, L and V (aliphatic), and S and T (tiny and polar). The latter amino acids are focused in the LG4X 'Slow' matrix, while the LG4M 'Slow' matrix mainly deals with tiny and nearly neutral amino acids (A, G, S and T) and the I, V pair (Sup. Mat.).

The 'Very Slow' matrices are thus used to express the fact that even in very slow sites, substitutions between highly similar amino acids are likely to occur. Their contents may be seen as being similar to that of the profiles in the CAT model (Lartillot and Philippe 2004; Le, Gascuel and Lartillot 2008). The 'Slow' matrices are analogous, but less contrasted. Moreover, the 'Slow' matrix of LG4M is relatively close to 'Buried' from EX3 (Tab. 1 and above hydropathy values), indicating (as expected) that buried sites and slow sites are often the same. However, both LG4M and LG4X 'Very Slow' matrices partly contradicts this basic fact, as the LG4X 'Very Slow' matrix focus on some hydrophilic pairs, and the LG4M 'Very Slow' matrix is slightly hydrophilic (Hydro = -0.429). An explanation of this finding could be that both LG4X and LG4M 'Very slow' matrices are strongly influenced by the genetic code (see above examples), which intervenes first in the mutational process (before the physicochemical constraints and selection) and favors substitutions between amino acids that are not necessarily hydrophobic.

The 'Medium' matrix of LG4M is correlated with both LG and 'Intermediate' matrix from EX3, and is thus mostly used for standard sites with average rates and solvent accessibility. The 'Medium' matrix of LG4X is also correlated with 'Intermediate', but to a lesser extent than LG4M 'Medium', and its global hydropathy is relatively low (-0.816) compared to that of LG (-0.253) and that of LG4M 'Medium' (0.219).

Lastly, the 'Fast' matrices of LG4X and LG4M are very close (correlation of the log-entries = 0.994) and quite similar to 'Exposed' from EX3 (correlation of the log-entries ≈ 0.99). As expected, fast sites and exposed sites are often the same. Moreover, 'Fast' matrices show a relatively low contrast (Fig. 2) and allow for all possible substitutions, with a preference for substitutions between amino acids with similar hydrophathy.

All together we thus see (as expected) a clear correlation between evolutionary rate and solvent accessibility; for example, LG4M 'Slow' is close to 'Buried' while both 'Fast' matrices are close to 'Exposed'. However, the matrices of LG4M and LG4X account for other features of the substitution processes; for example, LG4X 'Very Slow' is weakly correlated with 'Buried' but focuses on specific highly exchangeable amino acid pairs, some highly hydrophilic (e.g., R and K). The variability among all these replacement matrices demonstrates the complexity of amino acid substitutions and explains why a single matrix has limited capacity in modeling such complex processes.

Analysis of rates across sites further illustrates the difficulty involved in substitution modeling and the advantage of flexible models such as LG4X. With LG4M, the α value of the gamma parameter is significantly higher than with LG (respectively 0.866 and 0.584 on average; this ordering of α values is observed with all but five alignments, see Supp. Mat.). This is an expected outcome, as part of the rate variability is taken into account in LG4M by the use of four different matrices. With LG4X, the matrix rates show a different picture. While with LG4M each of the four matrices is pretty much associated with the same rate, with LG4X the rates differ significantly depending on the data set analyzed, to the point that the 'Slow' matrix is sometimes (3 cases among 84 TreeBase test alignments, Tab. 1) associated with the fastest rate. It is worth to note that that rates and weights in Equation (6) are optimized for each data set analyzed, without constraining the rates to be ordered depending on the matrix with which they are associated. In the same way, the weights of site categories are highly variable; for example with TreeBase test alignments, the 'Fast' weight varies from ~0.0 to ~0.25 (Tab. 1). Results with HSSP test alignments (Supp. Mat.) are similar but show less variability; for example, the 'Slow' matrix is only once the fastest one

among 300 alignments, instead of 3/84 with TreeBase. This indicates that the flexibility of LG4X is less useful with HSSP than with TreeBase, which can be expected since LG4X was estimated from HSSP.

The gains obtained by LG4X over LG4M (see next section) are thus explained by the high flexibility of LG4X, which is more powerful than the association of a single matrix with a distribution-free scheme of rates across sites, whose performance is somewhat disappointing (Susko et al. 2003; Mayrose, Friedman and Pupko 2005; see results of LG+Φ4 below). Here, each matrix corresponds (to some extent) to many/few and fast/slow sites, depending on the protein analyzed. This clearly shows that substitutions are not just Markovian with a fixed pattern (replacement matrix) modulated by site-dependent rates. As in other site-dependent models, we have different categories of sites corresponding to different matrices and tendencies to be slow or fast, but no strict constraint to be so. This further illustrates the finding by many authors (e.g., Keane et al. 2006) that substitution patterns may be very different in different proteins. The payoff of LG4X flexibility is that matrices in this model are not fully interpretable as 'Very Slow', 'Slow', 'Medium' and 'Fast' matrices; for example, the 'Slow' matrix is sometimes the fastest one and this feature most likely impacts its coefficient values. However, the average rates of these four matrices (0.29, 0.77, 1.37 and 3.42, respectively) clearly correspond to their natural interpretation. It must be emphasized that this ranking and global interpretation are obtained through our semi-supervised learning procedure (see above and Fig. 1), where only the first step accounts for site rates while further optimization steps are performed in a blind manner, clustering the sites based on their preferred matrix without reference to their rate. The fact that the four LG4X matrices are still clearly correlated to rate categories after this (6-step) phase of blind learning illustrates (if needed) that the evolutionary rate is a major factor in modeling substitution processes.

Performance comparisons

In this section, we assess the performance of the new models LG4M and LG4X by comparing them to existing models using 84 TreeBase and 300 HSSP test alignments (see Data Sets). The following models are compared:

Single-matrix models: JTT, WAG, LG. These standard matrices are used with four categories of gamma-distributed rates across sites (+Γ4 option, not indicated below for conciseness). To assess the use of a gamma distribution with four discrete rate categories, we ran: LG-Γ (constant site rate); LG+Γ3, LG+Γ6 and LG+Γ8 with 3, 6 and 8 gamma rate categories, respectively; LG+Φ4 (free distribution of site rates with 4 categories, just as in LG4X, but using a single LG matrix). Moreover, we tested LG+F, where the amino acid frequencies are estimated from the studied alignment, instead of being assigned to the default, average frequencies of LG. LG+F was used with the +Γ4 option, as LG and most other models studied here. LG+F should better fit the specificities of the data being analyzed, but is penalized by the large number of extra-parameters (frequencies) to be estimated. In total, LG+F has 20 free parameters (1 gamma + 19 frequencies); LG+Φ4 has 6 free parameters (3 rates + 3 weights); LG, JTT and LG (+Γ3, +Γ4, +Γ6, +Γ8) have 1 free (gamma) parameter; LG-Γ has 0 free parameter.

Two-level mixture models: EX2, UL3, EXEHO. EX2 involves two first-level categories of sites, based on solvent accessibility; its two matrices were estimated in a supervised manner from sites being classified as Buried and Exposed in HSSP. UL3 has three first-level categories of sites; it was estimated in a fully unsupervised manner, starting from 3 random matrices. EXEHO has six first-level categories of sites, crossing solvent accessibility (2 categories) with secondary structure (3 categories: Extended, Helix and Other); EXEHO was learned in a supervised manner from HSSP sites categorized accordingly. First-level categories in EX2, UL3 and EXEHO were combined in this study with four second-level gamma-distributed rate categories (+Γ4 option); for example, EXEHO has 6 x 4 = 24 site categories in total. EX2 and UL3 proved to be our best mixture models with two and three (first-level) categories, respectively, and UL3 was even better than the CAT60 model that

involves 60 first-level categories (profiles), and thus 240 categories in total with the +Γ4 option (Le, Lartillot and Gascuel 2008). Among mixture models, EX2 and UL3 were only beaten by EXEHO (Le and Gascuel 2010), but this latter requires high memory consumption and running time with large data sets due to its 24 site categories; for this reason, EXEHO was not run on TreeBase test alignments, some being very large, but on HSSP alignments only. We also tested the model by Wang et al. (2008), but encountered several difficulties when running their program and do not provide results for this model (e.g., QmmRAxML did not finish searching for 10/84 alignments after three weeks). EX2, UL3 and EXEHO require estimating the gamma rate parameter from the data plus the proportions of first-level categories, that is, 2, 3 and 6 free parameters, respectively.

Confidence-based models: EX2/S, EXEHO/S. The previous models are mixtures. EX2 and EXEHO use categories of sites having a structural meaning, and matrices estimated from sites categorized based on their structural properties, but to infer phylogenies these two models do not use any structural information. The likelihood of every site is computed within each category and then averaged, as expressed in Equation (4). On the contrary, EX2/S and EXEHO/S use structural information on the analyzed data set. Basically, the likelihood of each site is computed based on its known structural category, as in the standard partition approach. However, since structural information may be erroneous or inappropriate in a phylogenetic context, we refined this approach by introducing a confidence coefficient, estimated from the analyzed data set, which expresses a trade-off between the standard mixture (no structural information is available) and partition (structural information is fully reliable and relevant) models. These models are described in details in (Le and Gascuel 2010), where they are called EX2_CONF/MIX and EX_EHO_CONF/MIX. Here we use EX2/S and EXEHO/S to make it clear that they benefit (contrary to all other models) from structural information. Both are combined with four gamma rate categories (+Γ4 option). They were run on HSSP test alignments only, as no structural information is available in TreeBase. EX2/S and EXEHO/S involve 3 (1 gamma + 1 confidence + 1 category

proportion) and 6 (1 gamma + 1 confidence + 5 category proportions) free parameters to be estimated from the data.

Single-level mixture models: LG4M and LG4X. These are the two new models proposed in this paper. Both involve 4 site categories in total (to be compared to the 24 categories of EXEHO and the 240 of CAT60, remembering that the computing time and memory consumption are strongly correlated to the number of categories). Rates in LG4M are gamma distributed, while LG4X uses a distribution-free scheme. LG4M has 1 (gamma) free parameter; LG4X has 6 (3 rates + 3 weights) free parameters.

Comparison criteria and methods

Our aim was to compare the performance of all these models, regarding likelihood and topological criteria. To infer trees, we used: the last version of PhyML 3.0 (Guindon et al 2010) for LG, JTT and WAG; PhyML-Structure (Le and Gascuel 2010) for EX2, UL3, EXEHO, EX2/S and EXEHO/S; and our adaptation (PhyML-4X) of PhyML 3.0 for LG4M, LG4X and LG+Φ4. All programs were run with BioNJ (Gascuel 1997) starting tree and subtree pruning and regrafting (SPR) tree searching.

Since these models involve different numbers of free parameters, we measured their fitness to data using the AIC criterion (Akaike 1974)

$$AIC(M, D^a) = -2LL(M, T^a; D^a) + 2\#parameters(M),$$

where: $LL(M, T^a; D^a)$ is the log-likelihood of alignment D^a given model M and inferred tree T^a; $\#parameters(M)$ is the number of free parameters of model M. The AIC criterion has to be minimized; best scores are given to models with low numbers of free parameters and high likelihood values. All tested models involve one parameter (length) per tree branch, plus the model parameters detailed in previous section.

For every model M studied, we computed the average AIC per site for all alignments in test set A

$$AIC/site(M,A) = \frac{\sum_{a \in A} AIC(M, D^a)}{\sum_{a \in A} s^a},$$

where s^a is the number of sites in D^a. To complete this global average result, we performed pairwise model comparisons and counted the number of alignments D^a where $AIC(M_1, D^a) < AIC(M_2, D^a)$ (i.e., M_1 fits D^a better than M_2) for a given model pair M_1, M_2. To assess the statistical significance of the observed difference between M_1 and M_2 for any given alignment, we used a Kishino-Hasegawa (KH; 1989) test with p<0.01. As the number of free parameters between M_1 and M_2 may differ, we used AIC penalized likelihood values. This test is essentially the same as that used to compare phylogenies. We use the RELL bootstrap to estimate the distribution of the test statistic under the null hypothesis that both models are equivalent, but incorporate the number of parameters of each model in this statistic, just as in the AIC criterion (see Shimodaira 1997, for explanations and justifications of this test).

We compared the lengths of inferred trees, that is, the sums of their branch lengths. It has been suggested that best models tend to produce longer trees capturing more hidden substitutions (e.g., Pagel and Meade 2005).

We also compared the topologies of inferred trees. Indeed, if the new models produced the same topologies as the existing models, the effort of introducing new models would be rather useless. Unfortunately, the true tree is usually unknown with real data (as opposed to simulated data), and thus it is hard to assess the topological accuracy induced by any tree-building approach in a realistic setting. Here, we studied the topological impact of our new models, that is, whether or not using these models enables us to frequently infer trees that differ from those inferred with standard models.

When comparing models M_1 and M_2, we counted the number of alignments where the inferred topology using M_1 differs from that obtained using M_2. Both topologies were also compared using the Robinson and Foulds (RF; 1981) distance, which is the number of branches (bipartitions) that belong to one tree but not to the other. When different topologies

are found, one should prefer the one with best likelihood value, or best AIC (or similar criterion) value, when evolutionary models used for tree inference involve different numbers of parameters. However, the difference may be slight and non-significant, so one cannot reject the topology with a lower fit to data. We thus counted the number of alignments where M_1 and M_2 topologies differ, and where M_1 is significantly better (worse) than M_2, using a KH test on AIC penalized likelihood values with $p<0.01$.

Lastly, we checked that the observed topological differences comprised some branches with significant support. Indeed, the topological impact would be low if all differences corresponded to poorly-supported branches. To this end we performed bootstrap analyses and counted the number of branches with notable bootstrap support (BP1 $\geq$ 50%) in one tree, which were not recovered in the other tree, or had a much lower support in this tree (BP2 + 50% $\leq$ BP1). For example, one branch with BP1=40% was not counted, even when it was not recovered in the other tree; on the contrary, one branch with BP1=80% in one tree was counted when it was found in the other tree with BP2=20%. This measure (first introduced in Le and Gascuel 2010) thus summarizes the topological and branch support differences. We used only 50 bootstrap replicates for computing time reasons, but this suffices for our gap of 50% between BP1 and BP2 to be highly significant (p-value ~ 0.0 using a Z-test for two proportions). Moreover, 50% of bootstrap support was shown to be optimal in terms of topological error (Berry and Gascuel 1996; see also Holder, Sukumaran and Lewis 2008). As this procedure is computationally heavy (even with 50 replicates), we analyzed the 63 smallest TreeBase alignments only, and all (300, relatively small) HSSP alignments were analyzed.

Fitness comparison

Comparisons were performed on 84 TreeBase and 300 HSSP test alignments. Both sets are independent of the training alignments, and thus provide fair estimations of model performance. Moreover, using TreeBase should avoid possible biases induced by some of the specificities of HSSP alignments used to train our models. Tables 2 (TreeBase) and 3 (HSSP)

display comparisons between all models listed above. Figures 3 (TreeBase) and 4 (HSSP) show the progress in the AIC/site for the main models.

It is clear from these results that LG outperforms JTT and WAG. The average AIC/site of LG with TreeBase alignments is respectively 0.47 and 0.29 lower than that of JTT and WAG, equivalent to a gain of 117.5 and 72.5 log-likelihood units with a 500-site alignment. Moreover, LG is significantly better than JTT and WAG for most alignments. These results reconfirm the claim in (Le and Gascuel 2008).

Table 2 (TreeBase) reports comparisons between several standard model options, combined here with LG. The comparison between LG-Γ (no gamma distribution of site rates) and LG+Γ4 highlights the crucial role of modeling rates across sites. LG+Γ4 has significantly better AIC values than LG-Γ for 80/84 alignments, with an average AIC/site gain of 2.34. LG+Γ4 is significantly better than LG+Γ3 for 41/84 alignments, with an average AIC/site gain of 0.07, while it is worse than LG+Γ6 for 37/84 alignments, with a slight average AIC/site loss of 0.04. Using eight categories in LG+Γ8 does not improve AIC results, compared to LG+Γ6. These results indicate that with standard models three gamma rate categories are not enough, and that six categories suffice. Moreover, the differences in AIC/site between these options are low, compared to those of other options and models (e.g., LG+Γ versus LG-Γ). Using four categories, which is standard throughout the community, thus appears to be a fair compromise between likelihood (AIC) value and computing time, which is proportional to the number of categories. These results also support our choice of using four categories in our LG4M and LG4X models. Having three categories would be not enough and overly simple, while using four categories is likely to be a fair compromise, just as with standard models. However, we cannot exclude that having more than four categories could lead to even better (but slower) models. The low AIC/site gain (0.03) between LG+Φ4 and LG+Γ4 confirms the conclusions of Susko et al. (2003) and Mayrose, Friedman and Pupko (2005), who observed low gains when combining a single replacement matrix with a distribution-free scheme or a mixture of gamma distributions. Lastly, when combining LG with +F option, where amino acid distribution is estimated independently for each testing

alignment, the average AIC/site value is nearly the same as with LG alone. Six large alignments obtain significantly better AIC values than with LG, but for the other (small or medium-size) alignments the likelihood gain is not large enough to compensate for the 19 additional free parameters estimated from the data.

LG4M shows a clear improvement over LG, with an AIC/site gain of 0.15 and 0.59, with TreeBase and HSSP alignments respectively (note that these gains cannot be compared, as they depend on several factors, e.g. the number of taxa per alignment). With HSSP, LG4M has higher AIC (and likelihood) values than LG for 270/300 alignments with 174 significant cases. With TreeBase results are not so impressive, but still clearly in favor of LG4M versus LG.

LG4X has a major advantage over LG, with an AIC/site gain of 0.33 and 0.65, with TreeBase and HSSP respectively. LG4X is significantly better than LG for more than half of the alignments (both HSSP and TreeBase), while LG is significantly better than LG4X for only one (TreeBase) alignment. Compared to LG4M, LG4X shows a slight advantage with HSSP (AIC/site gain of 0.06) and is clearly better with TreeBase: mean AIC/site gain of 0.18, which is significant for 50/84 alignments, while LG4M is better than LG4X for only one non-significant alignment. The advantage of LG4X over LG4M is explained by its greater flexibility to fit the specificities of analyzed data, thanks to its distribution-free scheme. This scheme (+Φ4) does not show such an advantage with single-matrix models (see above), but becomes clearly beneficial when combined with different replacement matrices. The superiority of LG4X over LG4M is more marked with TreeBase than with HSSP alignments, possibly because LG4M and LG4X were learned from HSSP alignments and fit both HSSP specificities well. Moreover, HSSP alignments are smaller than TreeBase alignments, and some HSSP alignments are likely too small to compensate for the 5 additional parameters in LG4X, compared to LG4M. This advantage of LG4X over LG4M should thus be observed by future users analyzing phylogeny-intended alignments, such as those stored in TreeBase.

Compared with two-level mixture models, we see that LG4X is: (i) slightly better than EX2 (AIC/site gain of 0.08 and 0.15 with TreeBase and HSSP, respectively); (ii) nearly

equivalent to UL3 (AIC/site loss of 0.19 with TreeBase, but null with HSSP; the number of significant cases is low and does not favor one model over the other); (iii) slightly behind EXEHO (AIC/site loss of 0.14 with HSSP, but low number (23) of cases where EXEHO is significantly better than LG4X). Globally, we thus do not see a clear advantage of two-level mixture models over our new one-level mixture models, the former involving high number of rate categories (up to 24 with EXEHO) and heavy computational resources (at least EXEHO).

Lastly, when comparing EX2/S and EXEHO/S with LG4X and LG4M (and all other models), we see the clear advantage of using structure-informed models when the structural annotation of the proteins analyzed is available. The AIC/site gain of EXEHO/S over LG4X is of 0.61 on average, and this gain is significant with 223/300 alignments, while LG4X is never significantly better than EXEHO/S. The advantage of EX2/S over LG4X is less impressive, but still significant.

Tree-lengths comparisons (Tab. 2 and 3) do not show a clear picture. Some of the findings are expected, for example LG+Γ4 trees are much longer than LG-Γ ones, but the correlation between tree length and AIC value is weak or nonexistent. Mixture models tend to produce longer trees than standard models, with the notable exception of both distribution-free models (LG+Φ4 and LG4X) which infer trees shorter than LG+Γ4 ones. LG4M and LG+Γ4 trees have similar length. UL3 trees (comparable to LG4X in AIC terms) are very long, while those inferred using EXEHO/S (our best model in AIC terms) do not differ substantially from LG ones. These results thus contradict the assumption that better models should produce longer trees (Pagel and Meade 2005).

Overall, we the comparisons of likelihood and AIC values show that: (1) our new simple one-level mixture models outperform the standard models; (2) they are comparable to the best two-level mixture models, while requiring less computational resources; (3) they are clearly beaten by structure-informed models, which should be preferred when structural information is available.

Topological impact

The previous section analyzes model performance in terms of fit to the data, measured by likelihood and AIC values. Here we study the impact of using refined models, that is, how they change the topology of inferred trees.

We see from Tables 2 and 3 that refined models have a strong topological impact, compared to standard models. For example comparing LG4X to LG, both models infer different topologies with 58/84 TreeBase and 267/300 HSSP alignments. Moreover, these topologies are clearly different (percentage RF distance of 15% and 19% for TreeBase and HSSP, respectively), and the AIC value significantly favors LG4X topologies over LG topologies for 36/58 TreeBase and 163/267 HSSP alignments, while the LG topology is significantly favored for only one TreeBase alignment and never with HSSP. This means that with 36 (~45%) TreeBase and 163 (~55%) HSSP alignments, one should confidently select the LG4X topology and abandon that inferred using LG. Moreover, these two topologies are very different in general (see RF values).

However, the effective impact of selecting these alternative trees would be low if the differences between topologies of standard and refined models corresponded to poorly-supported branches and was solely due to random effects inherent to phylogenetic reconstruction. Indeed, inspecting the topological distances (RF) between LG topologies and those of other models, we see that closely-related models still show substantial topological differences; for example (Tab. 2), LG (used with the +Γ4 option, omitted below for conciseness as in the previous section) and LG+Γ8 topologies are different for 32/84 TreeBase alignments, with 9/32 significant cases and percentage RF distance of 7%. We thus counted the number of branches that are supported by the bootstrap in one tree but not the other (BP1≥BP2+50%, see above). For example, comparing LG and LG+Γ8 with the 63 smallest TreeBase alignments, we have 8 such branches among a grand total of 2,382 branches, against 48 with LG versus LG4X. Using the 300 HSSP alignments, we have 122 branches supported in one tree but not the other when comparing LG and LG+Γ8, and 552 for LG versus LG4X (grand total = 23,908). These measures are much lower than the

corresponding RF topological distances (Tab. 2 and 3), as expected since here we only consider branches with substantial bootstrap support. However, with LG versus LG4X we still have on average ~1 (TreeBase) to ~2 (HSSP) branches per tree with clearly different bootstrap support. Moreover, this "topological support dissimilarity" provides a sharper view of the models' resemblance and dissemblance, than standard RF distance; for example, the topological support dissimilarity between LG and LG4X is ~6 times larger with TreeBase (~4.5 with HSSP) than the topological support dissimilarity between LG and LG+Γ8, while this ratio is ~2 with RF distance (both TreeBase and HSSP).

We thus measured the topological support dissimilarities between main models using (63 smallest) TreeBase and (300) HSSP alignments, and then constructed distance-based trees representing the topological impacts and resemblance/dissemblance of these models. For this purpose, we used the FastME software (Desper and Gascuel 2003; http://www.atgc-montpellier.fr/fastme/) with default options (analogous to NJ but using branch swapping). Resulting trees are displayed in Figure 5, and the pairwise distance matrices are available in Supplementary Material.

While these two trees were obtained from completely different data (TreeBase, HSSP) through a complex procedure (bootstrap, dissimilarity computation, FastME), it is remarkable that both are nearly identical in terms of topology and branch lengths. Moreover, these trees are easily interpreted and illustrate the main features of studied models. A first obvious observation is that standard and non-standard models form the two main clades. Moreover, as expected with standard models: LG and LG+Γ8 form a tight clade; LG, WAG and JTT are relatively close; LG-Γ (constant site rate) is isolated at the end of a long branch, which illustrates the (well-documented) impact of using a gamma distribution of site rates. Non-standard models are separated in two clades respectively containing one-level and two-level mixtures. Within two-level mixtures in the HSSP tree we have a clade containing all structure-based models, with two tight sub-clades containing respectively EX2 and EX2/S, and EXEHO and EXEHO/S. Surprisingly, knowing the 3D structure in EX2/S and EXEHO/S does not have much impact on the tree topology with respect to EX2 and EXEHO. However,

the topological impact of these four structure-based models with respect to LG is almost the same as that of LG with respect to LG-Γ. The topological impact of UL3, LG4M and LG4X with respect to LG is even larger. As expected LG4M and LG4X form a clade, but both models are relatively distant, while UL3 is distant from all other models.

From trees in Figure 5, it is not possible to predict which model provides the best topologies. However, it is noticeable that non-standard and mixture models are on the opposite side of LG-Γ, known to be a poor model, while standard models are in between. We might have a preference for structure-based models, as they infer similar tree topologies (Fig. 5) and lead to very high likelihood values when the 3D structure is available (Fig. 4). However, there is no guaranty that these models are best from a topological standpoint, even if they have strong biological justifications. LG4X and LG4M are also based on meaningful assumptions (as opposed to UL3 learned in a purely blind manner). Both have a strong topological impact and provide high likelihood gains compared to standard models. LG4M and LG4X tree topologies contain well supported clades, not discovered by any of the other models, and thus representing biological and phylogenetic interest and deserving further investigations.

Memory consumption and running time

LG4M and LG4X require the same amount of memory as single-matrix models with the Γ4 option. EX2, UL3 and EXEHO (all three with the Γ4 option), respectively use twice, three and six times as much memory, because they are two-level mixtures with two, three and six matrices, respectively. For example, LG4X and UL3+Γ4 require two gigabytes (GB) and six GB of memory space, respectively, for the largest TreeBase alignment with 62 taxa and 11,544 sites (accession number M4680). EXEHO requires almost twelve GB to analyze this data set, which makes it impractical for most standard computers and users.

The same ratios apply to the computing times needed to calculate data likelihood, given a tree with branch lengths and model parameter values. For example, EXEHO is nearly six times slower than a standard, single-matrix model. However, other factors impact the total

computing time. Most notably, models differ in the number of parameters to be estimated from the data via likelihood optimization. Standard models and LG4M have only one (gamma) parameter, while EX2, UL3, LG4X and EXEHO respectively have two, three, six, and six parameters, thus again requiring additional computing time compared to standard models and LG4M.

Using PhyML-4X with standard options (+Γ4, SPR-based tree searching) on a powerful CPU (Intel(R) Xeon(R) E5440 at 2.83GHz with 16 GB memory) to infer phylogenies for all 84 TreeBase alignments, requires about 55 hours with standard models (LG was used here), 60 hours with LG4M, and 85 hours with LG4X. As expected, LG4M is nearly as fast as standard models, but LG4X is somewhat slowed down by model parameter estimations. Using PhyML-Structure (Le and Gascuel 2010) for the same task requires about 280, 380, and 670 hours for EX2, UL3 and EXEHO, respectively (total time for EXEHO is about one month and was estimated from a sample of alignments). Applying these models to the largest TreeBase alignment (M4860) using the same CPU, programs and options, requires about 6, 8, 11, 51, 53 and 84 hours for LG, LG4M, LG4X, EX2, UL3, and EXEHO, respectively. Though different programs were used in these experiments, with PhyML-4X being based on a more recent and about twice faster version of PhyML than PhyML-Structure, we obtain a clear picture: LG4M is nearly the same as standard models and is a fast model, as expected; LG4X is a bit slower due to its six parameters to be estimated; EX2 and UL3 are significantly slower than standard models, but still clearly applicable, even to large data sets; EXEHO requires important computing resources for large data sets, not only in terms of computing times but also with respect to memory space. The /S option (called CONF/MIX in Le and Gascuel, 2010) that we used for HSSP alignments with known 3D structure, requires nearly the same time and memory as the mixture version, with both EX2 and EXEHO. However, EXEHO (and EX2) may be used with less demanding options (CONF/LG and PART) when the 3D structure is known.

Conclusion

In this paper, we proposed two new models, LG4M and LG4X, for amino acid replacement modeling in protein phylogenetics. The main idea was to use different substitution matrices for the different evolutionary rate categories, and to reduce the standard gamma distribution constraints on site rates by adopting a distribution-free scheme (LG4X). Experiments with independent alignments showed that LG4M and LG4X most often infer trees with higher likelihood and AIC values than single-matrix models (JTT, WAG and LG), thus illustrating the limit of the standard approach that assumes a unique replacement matrix regardless of the site rate. Moreover, these trees tend to differ significantly in their topology from those inferred using the standard approach. These experiments also showed that our distribution-free scheme for site rates offers high flexibility and contributes greatly to LG4X performance.

Since LG4M and LG4X produce significantly better results while requiring the same memory space and similar running times, they would be reasonable replacements for single-matrix models. Current phylogenetic tree inference software using the standard approach and gamma distribution of site rates could immediately use LG4M because it requires the same data structures and procedures as single-matrix models. LG4X could be easily adapted as well, by adding an appropriate optimization procedure for estimating the distribution-free site rate parameters. These two models and others, notably structural, will be incorporated in a forthcoming release of the official PhyML.

Many questions arise from the improvement provided by LG4X and LG4M. The first is to check the number of rate categories. It is commonly acknowledged (our results in Tab. 2 support this practice) that four gamma-distributed rate categories provide a fair compromise for single-matrix models. Moreover, LG4X with four rate categories is better than two-matrix x four-rate models (EX2+Γ4) and comparable to three-matrix x four-rate models (UL3+Γ4). However, we do not yet know the difference when we increase or decrease the number of rate categories and use the same scheme as LG4M or LG4X. Variants of LG4X and LG4M and their combination with the standard +F option to fit proteins with specific amino acid

distributions should also be investigated. Lastly, a major direction for further research is to better understand the substitution processes revealed by these complex models and replacement matrices, unravel the biological differences between models and exploit them for further improvements.

Acknowledgements

Thanks to Associate Editor Jeff Thorne and two anonymous referees for their help and suggestions. This research was supported by the French ANR BIOSYS (MitoSys project) and the Vietnam National Foundation for Science and Technology Development.

References

Akaike H. 1974. A new look at statistical model identification. IEEE Transactions on Automatic Control AU-19: 716-722.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The protein data bank. Nucleic Acids Res. 28:235–242. Available from: http://www.pdb.org.

Berry V, Gascuel O. 1996. Interpretation of bootstrap trees : threeshold of clade selection and induced gain. Mol Biol Evol 13:999-1011.

Bryant D, Galtier N, Poursat MA. 2005. Likelihood calculations in phylogenetics. In: Gascuel O, editor. Mathematics of Evolution & Phylogeny. Oxford University Press, Oxford. p 33-62.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540-552.

Dayhoff MO, Eyck RV, Park CM. 1972. A model of evolutionary change in proteins. In: Dayhoff MO, editor. Atlas of protein sequence and structure. Volume 5. National Biomedical Research Foundation, Washington, DC. p 89-99.

Desper R, Gascuel O. 2002. Fast and Accurate Phylogeny Reconstruction Algorithms Based on the Minimum-Evolution Principle. J Comp Biol 19:687-705.

Durbin R, Eddy S, Krogh A, Mitchison G. 1998. Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge University Press.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368-376.

Felsenstein J. 2003. Inferring phylogenies. Sinauer Associates, Inc., Sunderland, MA.

Felsenstein J, Churchill GA. 1996. A Hidden Markov Model approach to variation among sites in rate of evolution. Mol Biol Evol. 13:93–104.

Gascuel O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol 14:685-695.

Gascuel O, Guindon S. 2007. Modelling the variability of evolutionary processes. In: Gascuel O, Steel M, editors. Reconstructing evolution: new mathematical and computational advances. Oxford University Press, Oxford. p 65-99.

Goldman N, Thorne JL, Jones DT. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics 149:445-458.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 59:307-321.

Holder MT, Sukumaran J, Lewis PO. 2008. A justification for reporting the majority-rule consensus tree in Bayesian phylogenetics. Syst Biol 57:814-821.

Holmes I, Rubin GM. 2002. An expectation maximization algorithm for training hidden substitution models. J Mol Biol. 3**17**:753-764.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci. 8:275-282.

Jones DT, Taylor WR, Thornton JM. 1994. A mutation data matrix for transmembrane proteins. FEBS Lett. 339:269-275.

Keane TMC, Creevey CJ, Pentony MM, Naughton TJ, McLnerney JO. 2006. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. BMC Evol Biol. 6:29.

Kishino H, Hasegawa M. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J Mol Evol. 29:170–179.

Klosterman PSA, Uzilov AV, Bendana YR, Bradley RK, Chao S, Kosiol C, Goldman N, Holmes I. 2006. XRate: a fast prototyping, training and annotation tool for phylo-grammars. BMC Bioinformatics 7:428.

Koshi JM, Goldstein RA. 1995. Context-dependent optimal substitution matrices. Protein Eng. 8:641-645.

Koshi JM, Goldstein RA. 1998. Models of natural mutations including site heterogeneity. Proteins 32:289-295.

Kyte J, Doolittle RF. 1982. A simple method for displaying the hydropathic character of a protein. J. Mol Biol. 157:105–132.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol. 21:1095-1109.

Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. Mol Biol Evol. 25:1307-1320.

Le SQ, Gascuel O, Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics 24:2317-2323.

Le SQ, Lartillot N, Gascuel O. 2008. Phylogenetic mixture models for proteins. Philos Trans R Soc Lond B. 363:3965–3976.

Le SQ, Gascuel O. 2010. Accounting for Accessibility to Solvent and Secondary Structure in Protein Phylogenetics is Clearly Beneficial. Syst Biol. 59:277–287.

Lio P, Goldman N, Thorne JL, Jones DT. 1998. PASSML: combining evolutionary inference and protein secondary structure prediction. Bioinformatics 14:726-733.

Mayrose I, Friedman N, Pupko T. 2005. A Gamma mixture model better accounts for among site rate heterogeneity. Bioinformatics 21:151-158.

Pagel M, Meade A. 2005. Mixture models in phylogenetic inference. In: Gascuel O, editor. Mathematics of Evolution & Phylogeny. Oxford University Press, Oxford. p 121-142.

Raman P, Cherezov V, Caffrey M. 2006. The Membrane Protein Data Bank. Cell Mol Life Sci. 63:36-51.

Robinson D, Foulds L. 1979. Comparison of weighted labeled trees. Lect. Notes Math. 748:119-126.

Sanderson MJ, Donoghue MJ, Piel W, Eriksson T. 1994. TreeBase: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. Amer Jour Bot. 81:183.

Schneider R, de Daruvar A, Sander C. 1997. The HSSP database of protein structure-sequence alignments. Nucleic Acids Res. 25:226-230.

Shimodaira H. 1997. Assessing the error probability of the model selection test. Ann Inst Stat Math. 49:395–410.

Susko E, Field C, Blouin C, Roger AJ. 2003. Estimation of rates-across-sites distributions in phylogenetic substitution models. Syst Biol. 52:594-603.

Thorne JL, Goldman N, Jones DT. 1996. Combining protein evolution and secondary structure. Mol Biol Evol. 13:666-673.

Wang HC, Li K, Susko E, Roger AJ. 2008. A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. BMC Evol Biol. 8:331.

Whelan S, 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. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol. 10:1396-1401.

Yang Z. 2006. Computational Molecular Evolution. Oxford University Press, Oxford.

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

TABLE AND FIGURE LEGENDS

Table 1. Main features of LG4M and LG4X replacement matrices

Note: The four matrices of LG4M and LG4X are ranked according to their average rates as 'Very Slow', 'Slow', 'Medium' and 'Fast'. 'LogCor/LG' is the Pearson correlation coefficient of the log-entries in the given matrix with those of LG. 'ClosestMatrix' is the matrix among 'Buried', 'Intermediate' and 'Exposed' matrices from EX3 (Le, Lartillot and Gascuel 2008) that is closest to the given matrix, based on the correlation of the log-entries (value in parentheses). 'Hydro' is the average hydropathy index (Kyte and Doolittle, 1982) of the 20 amino acids with weights equal to their equilibrium frequencies in the given matrix. 'Weight' is the weight (w) of the given matrix, averaged over the 84 TreeBase testing alignments. 'Rate' is the average rate (ρ) among TreeBase alignments. '5% quantiles' provide the 5th and 80th rate (resp. weight) values among these 84 alignments. 'Rate distribution' is the number of alignments in which a given matrix is ranked (based on its estimated rate) as very-slow/slow/medium/fast; for example, with 'Slow' we see that this matrix is never ranked as the slowest matrix, 75 times as the second slowest, 6 times as medium and 0 times as the fastest matrix. Similar statistics are obtained with the 300 HSSP test alignments (see Sup. Mat.).

Table 2: Model comparison with TreeBase test alignments, using likelihood and topological criteria

Note: Models are compared using 84 TreeBase test alignments. All models use four categories of gamma distributed rates (+Γ4), unless explicitly stated, that is: +Φ4 and LG4X for distribution-free scheme; -Γ for constant site rate; +Γ3 and +Γ8 for 3 and 8 gamma rate categories respectively. LG+F: LG exchangeability coefficients are combined with the amino frequencies of the alignment being analyzed. EX2: 2-matrix, two-level mixture model, with

matrices estimated from buried/exposed sites. UL3: 3-matrix, two-level mixture model, with blindly estimated matrices. LG4M: 4-matrix, one-level mixture model using gamma distribution of site rates, proposed in this paper. LG4X: 4-matrix, one-level mixture model using distribution-free scheme of site rates, proposed in this paper. On each row, model M1 is compared with model M2 using all test alignments. AIC/site: average per site difference in AIC value between M1 and M2; a positive (negative) value means that AIC/site of M1 is better (worse) than M2, on average. #M1>M2: number of alignments (out of 84) where M1 has a better AIC value than M2. #M1>M2 ($p<0.01$): number of alignments where the AIC of M1 is significantly better than that of M2; #M1<M2 ($p<0.01$): same as #M1>M2 ($p<0.01$), but now M2 is significantly better than M1. #T1>T2: number of alignments where the tree T1 inferred with M1 has a better AIC value than T2 inferred using M2 and where T1 and T2 have different topologies. #T1<T2: same as #T1>T2 but now T2 is better than T1. #T1>T2 ($p<0.01$): same as #T1>T2, but now T1 is significantly better than T2. #T2<T1 ($p<0.01$): T2 is significantly better than T1. RF (%): total Robinson and Foulds distance between T1 and T2 trees (i.e. sum over all data sets of the number of branches that belong to one tree but not the other); numbers in parentheses report the percentage of RF relative to the total number (3,994) of internal branches in both T1 and T2 trees. L1-L2 ($p<0.01$): average of tree length differences between T1 and T2; we also counted the number of cases where T1 is longer/shorter than T2 and assessed the significance using a sign test with $p<0.01$, significant differences are underlined.

Table 3: Model comparison with HSSP test alignments, using likelihood and topological criteria

Note: Models are compared using 300 HSSP test alignments. All models use four categories of gamma distributed rates (+Γ4), except LG4X. EX2: 2-matrix, two-level mixture model, with matrices estimated from buried/exposed sites. EX2/S has the same matrices as EX2, but

(contrary to EX2) uses the solvent accessibility of the residues, derived from the 3D protein structure. EXEHO: 6-matrix, two-level mixture model, combining accessibility to solvent and secondary structure. EXEHO/S has the same six matrices as EXEHO, but (contrary to EXEHO) uses the secondary structure and the solvent accessibility of the residues, derived from the 3D protein structure. UL3: 3-matrix, two-level mixture model, with blindly estimated matrices. LG4M: 4-matrix, one-level mixture model using gamma distribution of site rates, proposed in this paper. LG4X: 4-matrix, one-level mixture model using distribution-free scheme of site rates, proposed in this paper. On each row, model M1 is compared with model M2 using all test alignments. AIC/site: average per site difference in AIC value between M1 and M2; a positive (negative) value means that AIC/site of M1 is better (worse) than M2, on average. #M1>M2: number of alignments (out of 84) where M1 has a better AIC value than M2. #M1>M2 ($p<0.01$): number of alignments where the AIC of M1 is significantly better than that of M2; #M1<M2 ($p<0.01$): same as #M1>M2 ($p<0.01$), but now M2 is significantly better than M1. #T1>T2: number of alignments where the tree T1 inferred with M1 has a better AIC value than T2 inferred using M2 and where T1 and T2 have different topologies. #T1<T2: same as #T1>T2 but now T2 is better than T1. #T1>T2 ($p<0.01$): same as #T1>T2, but now T1 is significantly better than T2. #T2<T1 ($p<0.01$): T2 is significantly better than T1. RF (%): total Robinson and Foulds distance between T1 and T2 trees (i.e. sum over all data sets of the number of branches that belong to one tree but not the other); numbers in parentheses report the percentage of RF relative to the total number (23,908) of internal branches in both T1 and T2 trees. L1-L2 ($p<0.01$): average of tree length differences between T1 and T2; we also counted the number of cases where T1 is longer/shorter than T2 and assessed the significance using a sign test with $p<0.01$, significant differences are underlined.

Figure 1: Algorithm for estimating LG4M and LG4X

Note: Tree likelihood is calculated by PhyML-4X in step 3 using Equation (5) for LG4M and Equation (6) for LG4X. In step 6 $Q_k \approx Q_k^*$ is measured by the sum of squared entry differences, to be <0.1.

Figure 2: LG4M and LG4X replacement matrices

Note: Amino acids are ranked from highly hydrophilic (R) to highly hydrophobic (I), based on the hydropathy index (Kyte and Doolittle, 1982). Bubble sizes are proportional to the replacement rates. The horizontal axis displays the original amino acid, the vertical axis the new one resulting from replacement. X1: 'Very Slow' LG4X matrix; X4 'Fast' LG4X matrix (highly similar to fast LG4M matrix); M1: 'Very Slow' LG4M matrix; LG is provided as a reference average matrix.

Figure 3: AIC progress of amino acid replacement models, using TreeBase

Note: Models are compared using 84 TreeBase test alignments. All models use four categories of gamma distributed rates (+Γ4), except LG4X. EX2: 2-matrix, two-level mixture model, with matrices estimated from buried/exposed sites. UL3: 3-matrix, two-level mixture model, with blindly estimated matrices. LG4M: 4-matrix, one-level mixture model using gamma distribution of site rates, proposed in this paper. LG4X: 4-matrix, one-level mixture model using distribution-free scheme of site rates, proposed in this paper. In the upper panel (a) performance is measured by the average AIC per site (AIC/site) and compared with the JTT value. In the lower panel (b), we count the number of alignments (among 84) where each model provides a better (positive side) and a worse (negative side) likelihood value than LG. The black bars correspond to the numbers of significant differences using the Kishino-Hasegawa test on AIC values with p<0.01.

Figure 4: AIC progress of amino acid replacement models, using HSSP

Note: Models are compared using 300 HSSP test alignments. All models use four categories of gamma distributed rates (+Γ4), except LG4X. EX2: 2-matrix, two-level mixture model, with matrices estimated from buried/exposed sites. EX2/S has the same matrices as EX2, but (contrary to EX2) uses the solvent accessibility of the residues, derived from the 3D protein structure. EXEHO: 6-matrix, two-level mixture model, combining accessibility to solvent and secondary structure. EXEHO/S has the same six matrices as EXEHO, but (contrary to EXEHO) uses the secondary structure and the solvent accessibility of the residues, derived from the 3D protein structure. UL3: 3-matrix, two-level mixture model, with blindly estimated matrices. LG4M: 4-matrix, one-level mixture model using gamma distribution of site rates, proposed in this paper. LG4X: 4-matrix, one-level mixture model using distribution-free scheme of site rates, proposed in this paper. In the upper panel (a) performance is measured by the average AIC per site (AIC/site) and compared with the JTT value. In the lower panel (b), we count the number of alignments (among 300) where each model provides a better (positive side) and a worse (negative side) likelihood value than LG. The black and grey bars correspond to the numbers of significant differences using the Kishino-Hasegawa test on AIC values with $p<0.01$.

Figure 5: Topological support dissimilarities of the main models

Note: Topological support dissimilarity between models M1 and M2 is computed from the bootstrap trees inferred using M1 and M2, by counting the number of branches supported in one tree but not the other (BP1≥ BP2+50%, see text). Trees in this figure were built using distance-based FastME software, from all pairwise model dissimilarities. The tree in the upper panel (a) is based on the 63 smallest TreeBase test alignments; tree (b) is based on 300 HSSP test alignments. All models use four categories of gamma distributed rates (+Γ4),

unless explicitly stated, that is: LG4X for distribution-free scheme; -Γ for constant site rate; +Γ8 for 8 gamma rate categories. EX2: 2-matrix, two-level mixture model, with matrices estimated from buried/exposed sites. EX2/S has the same matrices as EX2, but (contrary to EX2) uses the solvent accessibility of the residues, derived from the 3D protein structure. EXEHO/S: 6-matrix model combining the accessibility to the solvent and the secondary structure of the residues, derived from the 3D protein structure. UL3: 3-matrix, two-level mixture model, with blindly estimated matrices. LG4M: 4-matrix, one-level mixture model using gamma distribution of site rates, proposed in this paper. LG4X: 4-matrix, one-level mixture model using distribution-free scheme of site rates, proposed in this paper.

Table 1. Main features of LG4M and LG4X replacement matrices

		Very Slow	Slow	Medium	Fast
LG4M	LogCor/LG	0.813	0.874	0.957	0.917
	ClosestMatrix	Intermediate (0.828)	Buried (0.914)	Intermediate (0.966)	Exposed (0.986)
	Hydro	-0.492	1.249	0.219	-1.682
	Average rate	0.145	0.440	0.952	2.463
	5% quantiles	0.035/0.315	0.257/0.673	0.826/1.053	1.938/2.882
LG4X	LogCor/LG	0.847	0.853	0.898	0.897
	ClosestMatrix	Buried (0.880)	Buried (0.885)	Intermediate (0.946)	Exposed (0.987)
	Hydro	0.934	0.325	-0.816	-1.815
	Average weight	0.313	0.332	0.233	0.122
	5% quantiles	0.180/0.418	0.185/0.419	0.145/0.375	0.019/0.251
	Average rate	0.289	0.770	1.370	3.420
	5% quantiles	0.084/0.441	0.394/1.209	0.800/2.232	1.406/5.317
	Rate distribution	84/0/0/0	0/75/6/3	0/9/73/2	0/0/5/79

Table 2: Model comparison with TreeBase test alignments, using likelihood and topological criteria

M1	M2	AIC/site	#M1>M2	#M1>M2 (p<.01)	#M1<M2 (p<.01)	#T1>T2	#T1<T2	#T1>T2 (p<.01)	#T1<T2 (p<.01)	RF (%)	L1-L2 (p<.01)
LG	JTT	0.47	73	66	7	39	8	36	5	430 (11)	0.036
LG	WAG	0.29	71	62	7	32	10	29	6	404 (10)	0.136
LG	LG-Γ	2.34	83	80	0	62	0	61	0	566 (14)	0.307
LG	LG+Γ3	0.07	80	41	1	32	0	11	0	242 (6)	0.002
LG	LG+Γ6	-0.04	7	1	37	3	22	0	6	266 (7)	0.018
LG	LG+Γ8	-0.05	10	1	33	3	29	0	9	270 (7)	0.027
LG	LG+Φ4	-0.03	34	15	21	17	28	4	10	476 (12)	0.063
LG	LG+F	0.00	51	7	6	24	12	4	3	364 (9)	-0.011
LG	LG4M	-0.15	33	11	34	20	37	7	24	616 (15)	-0.073
LG	LG4X	-0.33	12	1	50	10	48	1	36	606 (15)	0.062
LG4M	LG4X	-0.18	1	0	48	0	52	0	26	530 (13)	0.145
LG4X	EX2	0.08	67	27	2	44	11	17	1	536 (13)	-0.100
LG4X	UL3	-0.19	39	6	22	24	35	3	17	602 (15)	-0.265

Table 3: Model comparison with HSSP test alignments, using likelihood and topological criteria

M1	M2	AIC/site	#M1>M2	#M1>M2 (p<.01)	#M1<M2 (p<.01)	#T1>T2	#T1<T2	#T1>T2 (p<.01)	#T1<T2 (p<.01)	RF(%)	L1-L2 (p<.01)
LG	JTT	0.72	267	221	10	220	23	184	6	3570 (15)	0.048
LG	WAG	0.31	248	141	7	196	32	110	2	3478 (15)	0.174
LG	LG4M	-0.59	30	3	174	27	251	1	162	4386 (18)	0.009
LG	LG4X	-0.65	13	0	182	10	257	0	163	4548 (19)	0.078
LG4M	LG4X	-0.06	93	2	20	83	166	2	16	4014 (17)	0.068
LG4X	EX2	0.15	241	62	2	200	51	42	1	4110 (18)	-0.063
LG4X	UL3	0.00	199	37	10	165	99	26	10	4356 (18)	-0.326
LG4X	EXEHO	-0.14	117	2	23	88	166	2	21	4176 (17)	-0.094
LG4X	EX2/S	-0.21	60	1	80	56	199	0	57	4188 (18)	-0.058
LG4X	EXEHO/S	-0.61	5	0	223	4	250	0	181	4204 (18)	-0.092

Figure 1: Algorithm for estimating LG4M and LG4X

Estimation algorithm

step 1. Input a set of alignments, $\mathrm{D} = \{D^1, ..., D^N\}$

step 2. $\mathrm{Q} = \{Q_1 = Q_2 = Q_3 = Q_4 = LG\}$
; Initialization of the four matrices of* Q *with LG ;

step 3. For each alignment D^a

- Use PhyML-4X to estimate the tree T^a, rates $\rho^a = \rho_1^a, ..., \rho_4^a$, and weights $\mathrm{w}^a = w_1^a, ..., w_4^a$, based on Equation (8).
  ; ML estimation of the phylogeny and rate parameters for every alignment ;
- Cluster every site D_i^a of D^a into set $C_{c_i}^a$ where
  $c_i = \arg\max_{k=1..4} w_k L(T^a, \rho_k^a Q_k \mid D_i^a)$
  ; Estimation of the MAP site rate category ;

step 4. Build 4 sub-alignments from D^a with corresponding trees $\left(C_1^a, T^a \times \rho_1^a\right)$, $\left(C_2^a, T^a \times \rho_2^a\right)$, $\left(C_3^a, T^a \times \rho_3^a\right)$, and $\left(C_4^a, T^a \times \rho_4^a\right)$
; T* x *ρ means that all branch lengths in T are multiplied by ρ ;
; This is equivalent to multiplying the Q matrix in Equations (5), (6) by ρ;

step 5. For $k = 1..4,$ estimate Q_k^* using XRate with Equation (11) from sub-alignments $\left\{(C_k^1, T^1 \times \rho_k^1), ..., (C_k^N, T^N \times \rho_k^N)\right\}$
; Most standard use of XRate ;
; Thanks to tree scaling in step 4, branch lengths are comparable among ;
; sub-alignments, and the estimated matrix is expected to be (nearly) normalized ;

step 6. If $Q_k \approx Q_k^* : \forall k = 1..4$, then output $\mathrm{Q}^* = \left(Q_1^*, Q_2^*, Q_3^*, Q_4^*\right)$. Otherwise, replace $Q_k = Q_k^* : k = 1..4$ and go back to step 3.
; When convergence is observed, we output the estimated, normalized matrix ;
; Otherwise we iterate the computations ;

Figure 2: LG4M and LG4X replacement matrices

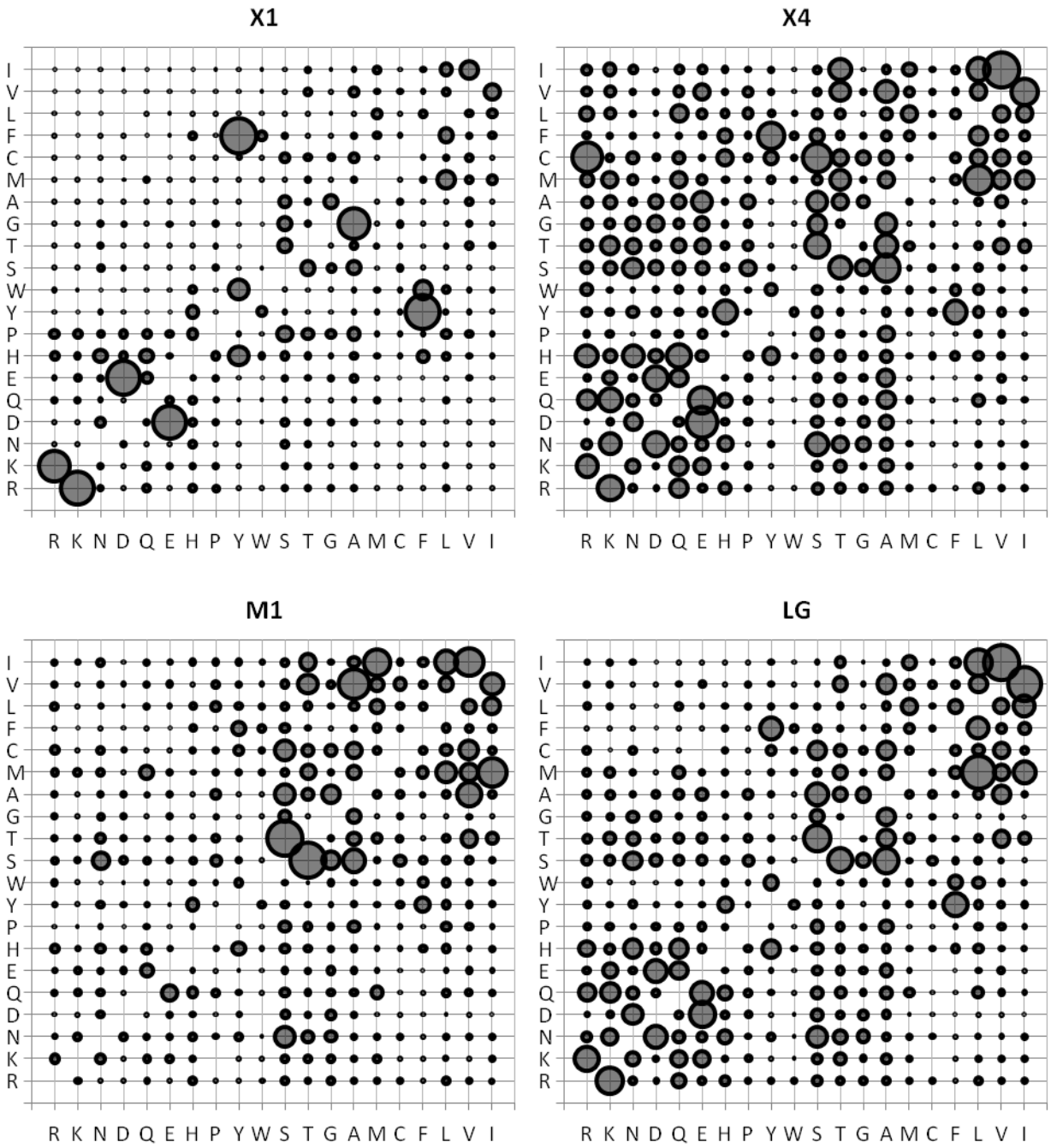

Figure 3: AIC progress of amino acid replacement models, using TreeBase

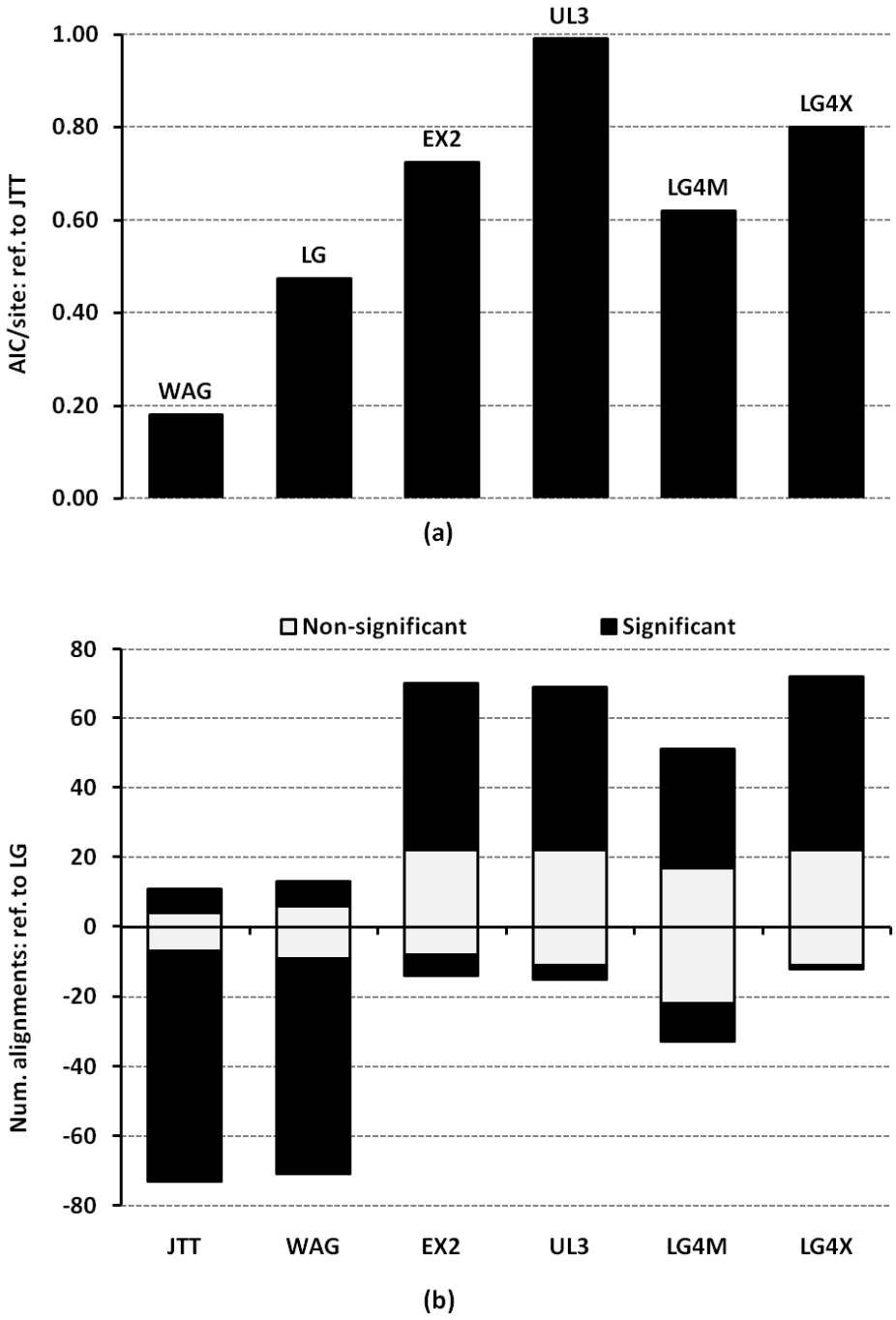

Figure 4: AIC progress of amino acid replacement models, using HSSP

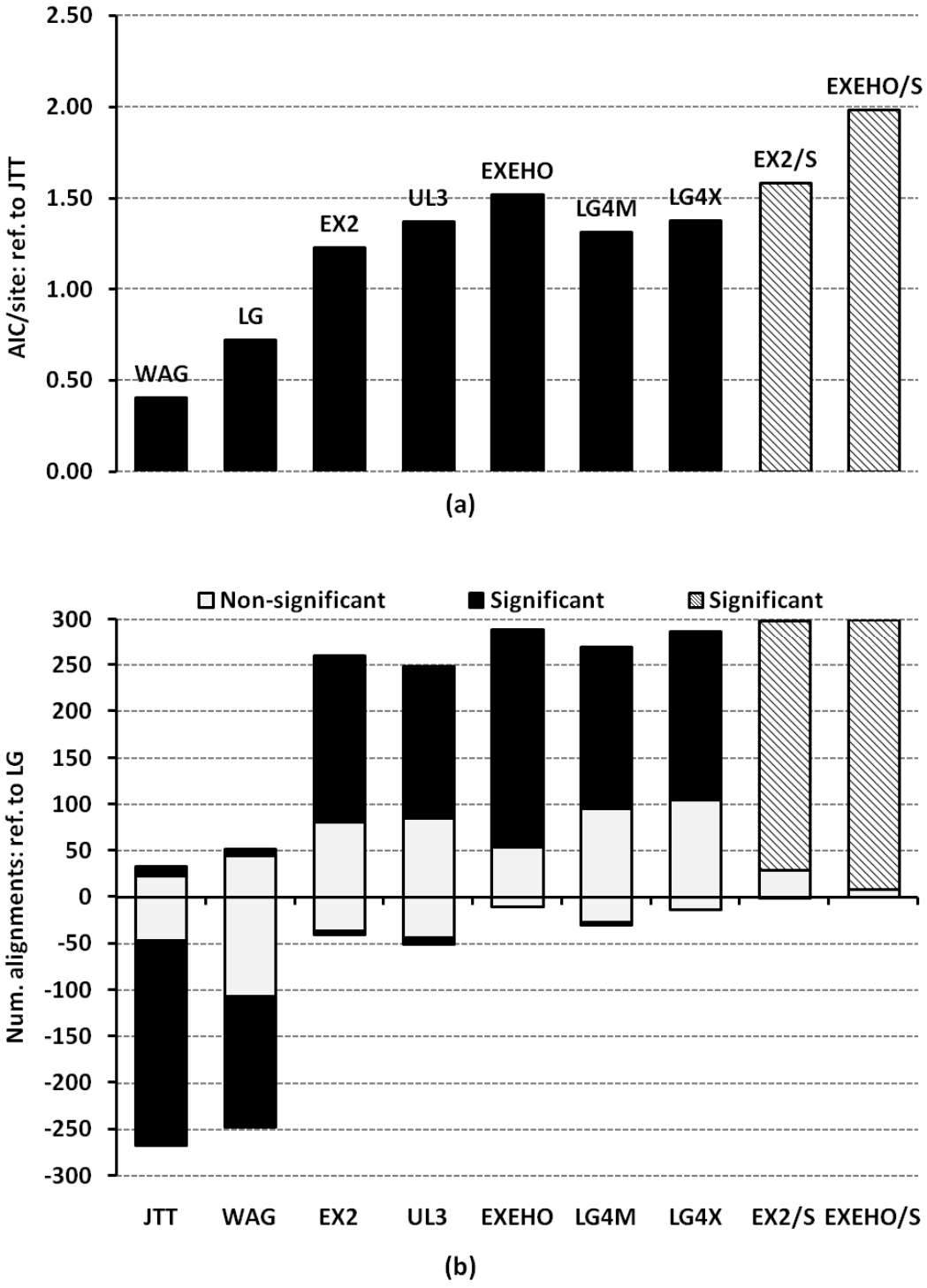

Figure 5: Topological support dissimilarities of the main models