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

1Méthodes et Algorithmes pour la Bioinformatique (LIRMM & IBC), Centre National de la Recherche Scientifique (CNRS)–Université Montpellier II, Montpellier Cedex 5, France
2Wellcome Trust Sanger Institute, Genome Campus, Hinxton, United Kingdom
3University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam

*Corresponding author: E-mail: gascuel@lirmm.fr.
Associate editor: Jeffrey Thorne

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 (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 with 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.atgc-montpellier.fr/models/lg4x.

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

Introduction

Amino acid replacement matrices—20 × 20 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 physicochemical 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 (ML) 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 et al. (1972), notably JTT (Jones et al. 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 et al. 1994) or mitochondrial (Yang et al. 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, whereas 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 et al. 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 et al. 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—alogous to the (Felsenstein 1981) model of DNA substitution—and do not use replacement matrices (or only highly simplified ones). In the multimatrix approach (Koshi and Goldstein 1995; Thorne et al. 1996; Goldman et al. 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 (for a similar model, see also Lartillot and Philippe 2004). In all cases, the set of profiles or matrices is combined thanks to a mixture approach or a Hidden Markov Model (HMM; Felsenstein and Churchill 1996; Thorne et al. 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 et al. 2005) to the number of site categories (e.g., 12 with 4 gamma categories and 3 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 one of the 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) and 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 the Homology-Derived Secondary Structure of Proteins database (HSSP; Schneider et al. 1997). HSSP comprises ~50,000 alignments of protein families, each containing an average of ~500 members. Each alignment is obtained by aligning a protein with known 3D 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 with an average of 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 PDB; Raman et al. 2006) since their amino acid replacement pattern is highly different to that of globular proteins (Jones et al. 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 nonredundant 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 Larillot 2008; Le, Larillot, 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 http://www.atgc-montpellier.fr/models/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 http://www.atgc-montpellier.fr/models/lg4x.

Models

All amino acid substitution matrices discussed here comply with the general time-reversible model (see textbooks, e.g., Bryant et al. 2005; Yang 2006). Such a matrix is denoted \( Q = (q_{xy}) \), 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 \( R = (r_{x \leftrightarrow y}) \) and the amino acid equilibrium distribution \( \pi = (\pi_x) \), using \( q_{xy} = \pi_x r_{x \leftrightarrow y} (x \neq y) \). The amino acid distribution \( \pi \) 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 \( \rho \) is the global rate of \( Q \). This constraint is released with some multimatrix models (e.g., Le, Larillot, and Gascuel 2008), where some site categories and matrices are fast with a high global rate and some others are slow with a low \( \rho \) 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 \( \pi - 1 \) normalization constraint).

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

\[
L(T, Q; D) = \prod_i L(T, Q; D_i),
\]

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 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

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

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

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

\[
L(T, Q = \{Q_1, \ldots, Q_M\}, w = \{w_1, \ldots, w_M\}; D) = \prod_i \left( \sum_{m=1}^{M} w_m L(T, Q_m; D_i) \right),
\]

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

Recent works (e.g., Le, Larillot, and Gascuel 2008; Wang et al. 2008) combined Yang’s model (2) with the above (3) multi-matrix model:

\[
L(T, Q = \{Q_1, \ldots, Q_M\}, w = \{w_1, \ldots, w_M\}, \alpha; D) = \prod_i \left( \sum_{m=1}^{M} w_m \sum_{k=1}^{K} L(T, \Gamma(\alpha, k); Q_m; D_i) \right),
\]

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, Larillot, 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 \( \pi \).

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, 12 with UL3 and 4 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 \( Q = \{Q_1, Q_2, Q_3, Q_4\} \) be the set of these four matrices, where \( Q_i \) stands for the...
matrix corresponding to slowest category and $Q_4$ that of the fastest one. Data likelihood is expressed as

$$L(T, Q, \alpha; D) = \prod_i \left( \frac{1}{4} \sum_k^4 L[T, \Gamma(\alpha, k)Q_k; D_1] \right).$$  \hfill (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., $\alpha$ 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 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:

$$L(T, Q, \rho = \{\rho_1, \rho_2, \rho_3, \rho_4\}; w = \{w_1, w_2, w_3, w_4\}; D) = \prod_i \left( \sum_k^4 w_k L[T, \rho_k Q_k; D_1] \right),$$ \hfill (6)

where $w_k$ and $\rho_k$ 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 $\rho_k$, which are estimated by maximizing likelihood (6) on the data set analyzed.

### Model Estimation

We have a set of $N$ protein alignments denoted $D = \{D_1, \ldots, D_N\}$, where $D^a$ is an alignment. We aim to estimate a 4-matrix model $Q^* = (Q_1^*, Q_2^*, Q_3^*, Q_4^*)$ that maximizes the likelihood of $D$:

$$Q^* = \arg \max_{Q = (Q_1, Q_2, Q_3, Q_4)} \left\{ \prod_i^N L(T, Q, \rho^a, w^a; D^a) \right\},$$ \hfill (7)

where $T = (T_1, \ldots, T_N)$, $P = (p^1, \ldots, p^N)$, and $W = (w_1, \ldots, w_N)$ are the trees, and weights of the $N$ alignments, respectively; $L(T, Q, \rho^a, w^a; D^a)$ is the likelihood of $D^a$ given model $Q$, tree $T^a$, rates $\rho^a = (\rho_1^a, \ldots, \rho_4^a)$, and weights $w^a = (w_1^a, \ldots, w_4^a)$. Thus, to estimate $Q^* = (Q_1^*, Q_2^*, Q_3^*, Q_4^*)$, we also need to estimate $T^a$, $P^a$, and $W^a$, which optimize likelihood (7). We are interested in two 4-matrix models: LG4M where $L(T, Q, \rho^a, w^a; D^a)$ is calculated using equation (5) and LG4X that is based on equation (6). For each alignment $D^a$, $\rho^a$, and $w^a$ of LG4M follow a discrete gamma distribution with four equally weighted rate categories, whereas in LG4X, these parameters are freely estimated such that $\sum_{k=1}^4 w_k^a = 1$ and $\sum_{k=1}^4 w_k^a \rho_k^a = 1$, without any additional constraint.

Following (Whelan and Goldman 2001; Le and Gascuel 2008; Le, Lartillot, and Gascuel 2008), we estimate $Q^* = (Q_1^*, Q_2^*, Q_3^*, Q_4^*)$ from equation (7) in two steps: 1) given a fixed starting value of $Q$, we estimate $T^a$, $P^a$, and $W^a$ by maximizing likelihood (5) (LG4M) or (6) (LG4X); 2) we then estimate $Q^* = (Q_1^*, Q_2^*, Q_3^*, Q_4^*)$ using equation (7) with respect to $T^a$, $P^a$, and $W^a$ values obtained in step (1). These two steps are iterated one after the other until no more improvement of $T^a$, $P^a$, $W^a$, and $Q^*$ is found.

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

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

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

$$Q^* = (Q_1^*, Q_2^*, Q_3^*, Q_4^*) = \arg \max_{Q = (Q_1, Q_2, Q_3, Q_4)} \left\{ \prod_i^N L(T, Q, \rho^a, P^a, W^a; D^a) \right\}. \hfill (9)$$

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

$$Q^* = (Q_1^*, Q_2^*, Q_3^*, Q_4^*) = \arg \max_{Q = (Q_1, Q_2, Q_3, Q_4)} \left\{ \prod_i^N \prod_{c_i = k} L(T, p^a, c_i; D^a) \right\}, \hfill (10)$$

where $D^a_i$ is the $i$th site of alignment $D^a$, $c_i$ is the MAP rate category (computed during tree estimation) for site $D^a_i$, and $p^a$, is the rate of $c_i$ corresponding to $Q_k$, substitution matrix. Equation (10) can then be rewritten as

$$\forall k = 1 \ldots 4, \quad Q^*_k = \arg \max_{Q_k} \left\{ \prod_i \prod_{c_i = k} L(T, p^a, c_i; D^a_i) \right\}. \hfill (11)$$

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) and 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 to estimate
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 cor-
rect matrix for each rate category. The other parts (e.g., to
optimize \( \alpha \) 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 \((w_1, \ldots, w_4)\) and rates \((\rho_1, \ldots, \rho_4)\),
alternating monodimensional Brent optimization of every
variable until global convergence. To account for constraint
\( \sum w_m = 1 \), we use the variable change \( w_i = e^{\nu_i} / \sum e^{\nu_i} \) and
then optimize the \( \nu_i \)'s 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 starting tree, rates, and weights (\( \approx 0.25 \))
are first estimated with LG4M, which involves a single
\( \alpha \) 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 semisupervised way. During the first step,
sites are categorized based on the rate (associated to
LG) providing the highest likelihood value. During sub-
sequent 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

Estimation algorithm

1. Input a set of alignments, \( D = \{D^1, \ldots, D^N\} \)
2. \( Q = \{Q_1 = Q_2 = Q_3 = Q_4 = LG\} \)
   \( ; \) Initialization of the four matrices of \( Q \) with \( LG \);
3. For each alignment \( D^a \)
   - Use PhyML-4X to estimate the tree \( T^a \), rates \( \rho^a = \rho^a_1, \ldots, \rho^a_4 \), and weights
     \( w^a = w^a_1, \ldots, w^a_4 \), based on Equation (8).
     \( ; \) ML estimation of the phylogeny and rate parameters for every alignment ;
   - Cluster every site \( D^a_i \) of \( D^a \) into set \( C^a_i \) where
     \( c_i = \arg \max_{k = 1, 4} w_k L(T^a, \rho^a_k | D^a) \)
   \( ; \) Estimation of the MAP site rate category ;
4. Build 4 sub-alignments from \( D^a \) with corresponding trees
   \( (C^a_1, T^a \times \rho^a_1), (C^a_2, T^a \times \rho^a_2), (C^a_3, T^a \times \rho^a_3), \) and \( (C^a_4, T^a \times \rho^a_4) \)
   \( ; T \times \rho \) means that all branch lengths in \( T \) are multiplied by \( \rho \);
   \( ; \) This is equivalent to multiplying the \( Q \) matrix in Equations (5), (6) by \( \rho \);
5. For \( k = 1, 4 \), estimate \( Q^*_k \) using XRate with Equation (11)
   from sub-alignments \( \{(C^k_1, T^k \times \rho^k_1), \ldots, (C^k_4, T^k \times \rho^k_4)\} \)
   \( ; \) 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 ;
6. If \( Q_k = Q^*_k : \forall k = 1, 4 \), then output \( Q^* = \{Q^*_1, Q^*_2, Q^*_3, Q^*_4\} \). 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 ;

Fig. 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 \( < 0.1 \).
(and weights) are estimated independently for each alignment without any a priori constraint, it may occur for some data sets that the “Fast” matrix is actually associated with a slow rate and vice versa (see Table 1 and supplementary tables and figures at http://www.atgc-montpellier.fr/models/lg4x). Thus, this semisupervised 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 are (mostly) preserved along all training steps, this implies that the rate factor is of first importance, as we assumed in this study.

As all Expectation–Maximization (EM) approaches, XRater 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 with other models), we did not try alternative starting matrices and training strategies. However, based on our previous experiments with LG where XRater 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 semisupervised 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 at http://www.atgc-montpellier.fr/models/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; Table 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 index (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 those 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

### Table 1. Main Features of LG4M and LG4X Replacement Matrices.

<table>
<thead>
<tr>
<th></th>
<th>Very Slow</th>
<th>Slow</th>
<th>Medium</th>
<th>Fast</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LG4M</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LogCor/LG</td>
<td>0.813</td>
<td>0.874</td>
<td>0.957</td>
<td>0.917</td>
</tr>
<tr>
<td>ClosestMatrix</td>
<td>Intermediate (0.828)</td>
<td>Buried (0.914)</td>
<td>Intermediate (0.966)</td>
<td>Exposed (0.986)</td>
</tr>
<tr>
<td>Hydro</td>
<td>−0.492</td>
<td>1.249</td>
<td>0.219</td>
<td>−1.682</td>
</tr>
<tr>
<td>Average rate</td>
<td>0.145</td>
<td>0.440</td>
<td>0.952</td>
<td>2.463</td>
</tr>
<tr>
<td>5% quantiles</td>
<td>0.035/0.315</td>
<td>0.257/0.673</td>
<td>0.826/1.053</td>
<td>1.938/2.882</td>
</tr>
<tr>
<td><strong>LG4X</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LogCor/LG</td>
<td>0.847</td>
<td>0.853</td>
<td>0.898</td>
<td>0.897</td>
</tr>
<tr>
<td>ClosestMatrix</td>
<td>Buried (0.880)</td>
<td>Buried (0.885)</td>
<td>Intermediate (0.946)</td>
<td>Exposed (0.987)</td>
</tr>
<tr>
<td>Hydro</td>
<td>0.934</td>
<td>0.325</td>
<td>−0.816</td>
<td>−1.815</td>
</tr>
<tr>
<td>Average weight</td>
<td>0.313</td>
<td>0.332</td>
<td>0.233</td>
<td>0.122</td>
</tr>
<tr>
<td>5% quantiles</td>
<td>0.180/0.418</td>
<td>0.185/0.419</td>
<td>0.145/0.375</td>
<td>0.019/0.251</td>
</tr>
<tr>
<td>Average rate</td>
<td>0.289</td>
<td>0.779</td>
<td>1.370</td>
<td>3.420</td>
</tr>
<tr>
<td>5% quantiles</td>
<td>0.086/0.441</td>
<td>0.396/1.209</td>
<td>0.800/2.232</td>
<td>1.406/5.317</td>
</tr>
<tr>
<td>Rate distribution</td>
<td>84/0/0/0</td>
<td>0/75/6/3</td>
<td>0/9/73/2</td>
<td>0/0/5/79</td>
</tr>
</tbody>
</table>

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 (μ) among TreeBase alignments. “5% quantiles” provide the 5th and 80th rate and 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 3 times as the fastest matrix. Similar statistics are obtained with the 300 HSSP test alignments (see supplementary tables and figures at http://www.atgc-montpellier.fr/models/lg4x).
be noted that rates displayed in figure 2 are relative rates; all matrices are normalized and do not incorporate the fact that very slow sites globally evolve slower (~6 times on average; table 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, whereas the LG4M “Slow” matrix mainly deals with tiny and nearly neutral amino acids (A, G, S, and T) and the I, V pair (supplementary tables and figures at http://www.atgc-montpellier.fr/models/lg4x).

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 (table 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 contradict this basic fact, as the LG4X “Very Slow” matrix focuses 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 hydrophathy is relatively low (−0.816) compared with 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", whereas 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 z value of the gamma parameter is significantly higher than with LG (respectively 0.866 and 0.584 on average; this ordering of z values is observed with all but five alignments, see supplementary tables and figures at http://www.atgc-montpellier.fr/models/lg4x. 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 (three cases among 84 TreeBase test alignments; table 1) associated with the fastest rate. It is worth to note 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 (table 1). Results with HSSP test alignments (supplementary tables and figures at http://www.atgc-montpellier.fr/models/lg4x) are similar but show less variability; for example, the “Slow” matrix is only once the fastest one among 300 alignments, instead of 3/8 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 Performance Comparisons) 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 et al. 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 semisupervised 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.

Model Comparisons

In the following, we assess the performance of the new models LG4M and LG4X by comparing them with 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 (+Φ4 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 four 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. LG+F should better fit the specificities of the data being analyzed, but is penalized by the large number of extraparameters (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, +Γ10, +Γ12).
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 three random matrices. EXEHO has six first-level categories of sites, crossing solvent accessibility (two categories) with secondary structure (three 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 × 4 = 24 site categories in total. EX2 and UL3 proved to be our best mixture models with 2 and 3 (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 3 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 with 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, whereas 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):

\[
\text{AIC}(M, D^a) = -2 \text{LL}(M, T^a; D^a) + 2# \text{parameters}(M),
\]

where \(\text{LL}(M, T^a; D^a)\) 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 \(A\):

\[
\text{AIC/site}(M, A) = \frac{\sum_{a \in A} \text{AIC}(M, D^a)}{\# \text{sites}(D^a)},
\]

where \#sites 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(M1,D^a) < AIC(M2,D^a) (i.e., \(M_1\) fits \(D^a\) better than \(M_2\)) for a given model pair \(M_1, M_2\). To test 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
Table 2. Model Comparison with TreeBase Test Alignments, Using Likelihood and Topological Criteria.

<table>
<thead>
<tr>
<th>M1</th>
<th>M2</th>
<th>AIC/site</th>
<th>#M1 &gt; M2 (P &lt; 0.01)</th>
<th>#M1 &lt; M2 (P &lt; 0.01)</th>
<th>#T1 &gt; T2 (P &lt; 0.01)</th>
<th>#T1 &lt; T2 (P &lt; 0.01)</th>
<th>RF (%)</th>
<th>L1 − L2 (P &lt; 0.01)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LG</td>
<td>JTT</td>
<td>0.47</td>
<td>73</td>
<td>66</td>
<td>7</td>
<td>39</td>
<td>5</td>
<td>430 (11)</td>
</tr>
<tr>
<td>LG</td>
<td>WAG</td>
<td>0.29</td>
<td>71</td>
<td>62</td>
<td>7</td>
<td>32</td>
<td>10</td>
<td>29</td>
</tr>
<tr>
<td>LG</td>
<td>LG−Γ</td>
<td>2.31</td>
<td>83</td>
<td>80</td>
<td>0</td>
<td>62</td>
<td>0</td>
<td>61</td>
</tr>
<tr>
<td>LG</td>
<td>LG+Γ3</td>
<td>0.07</td>
<td>80</td>
<td>41</td>
<td>1</td>
<td>32</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>LG</td>
<td>LG+Γ6</td>
<td>−0.04</td>
<td>7</td>
<td>1</td>
<td>37</td>
<td>3</td>
<td>22</td>
<td>0</td>
</tr>
<tr>
<td>LG</td>
<td>LG+Γ8</td>
<td>−0.05</td>
<td>10</td>
<td>1</td>
<td>33</td>
<td>3</td>
<td>29</td>
<td>0</td>
</tr>
<tr>
<td>LG</td>
<td>LG+Φ4</td>
<td>−0.03</td>
<td>34</td>
<td>15</td>
<td>21</td>
<td>17</td>
<td>28</td>
<td>4</td>
</tr>
<tr>
<td>LG</td>
<td>LG+F</td>
<td>0.00</td>
<td>51</td>
<td>7</td>
<td>6</td>
<td>24</td>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>LG</td>
<td>LG4M</td>
<td>−0.15</td>
<td>33</td>
<td>11</td>
<td>34</td>
<td>20</td>
<td>37</td>
<td>7</td>
</tr>
<tr>
<td>LG</td>
<td>LG4X</td>
<td>−0.33</td>
<td>12</td>
<td>1</td>
<td>50</td>
<td>10</td>
<td>48</td>
<td>1</td>
</tr>
<tr>
<td>LG4M</td>
<td>LG4X</td>
<td>−0.18</td>
<td>1</td>
<td>0</td>
<td>48</td>
<td>0</td>
<td>52</td>
<td>0</td>
</tr>
<tr>
<td>LG4X</td>
<td>EX2</td>
<td>0.08</td>
<td>67</td>
<td>27</td>
<td>2</td>
<td>44</td>
<td>11</td>
<td>17</td>
</tr>
<tr>
<td>LG4X</td>
<td>UL3</td>
<td>−0.19</td>
<td>39</td>
<td>6</td>
<td>22</td>
<td>24</td>
<td>35</td>
<td>3</td>
</tr>
</tbody>
</table>

Note.—Models are compared using 84 TreeBase test alignments. All models use four categories of gamma distributed rates (+Γ), unless explicitly stated, that is: +Φ4 and LG4X for distribution-free scheme; −Γ for constant site rate; −Γ+Γ, −Γ+Φ, and −Γ+Φ for 3, 6, and 8 gamma rate categories, respectively. LG−Γ: 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 (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 M12 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. RF (%): total RF 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 underline.

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 M1 and M2, we counted the number of alignments where the inferred topology using M1 differs from that obtained using M2. Both topologies were also compared using the Robinson and Foulds (RF; 1979) distance, which is the number of branch (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 nonsignificant, 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 \( P \leq 0.01 \).

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 \( \approx 0.0 \) 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 et al. 2008). As this procedure is computationally heavy (even with 50 replicates), we analyzed the 63 smallest TreeBase alignments only. 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, whereas 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 with 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 with those of other options and models (e.g., LG+Γ vs. 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, whereas 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 et al. (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 shows 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), whereas LG is significantly better than LG4X for only one (TreeBase) alignment. Compared with 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, whereas LG4M is better than LG4X for only one nonsignificant 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 with 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 1) slightly better than EX2 (AIC/site gain of 0.08 and 0.15 with TreeBase and HSSP, respectively); 2) 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); 3) 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, whereas LG4X is never significantly better than LG4X. Moreover, HSSP alignments are smaller than TreeBase alignments, possibly because LG4M and LG4X were learned from TreeBase.

Overall, 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; and 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 with standard models. For example comparing LG4X with 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, whereas 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 refined and standard 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 (table 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 (tables 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 dissimilarity 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

![Figure 5](https://academic.oup.com/mbe/article-abstract/29/10/2921/1027701/fig5.png)
(300) HSSP alignments, and then constructed distance-based trees representing the topological impacts and resemblance/dissimilarity of these models. For this purpose, we used the FastME software (Desper and Gascuel 2002; 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 the supplementary tables and figures at http://www.atgc-montpellier.fr/models/lg4x.

Although 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 nonstandard models form the two main clades. Moreover, as expected with standard models: LG and LG+$\Gamma$ form a tight clade; LG, WAG, and JTT are relatively close; LG-$\Gamma$ (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. Nonstandard 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 subclades containing, respectively, EX2 and EX2/S and EXEH and EXEH/S. Surprisingly, knowing the 3D structure in EX2/S and EXEH/S does not have much impact on the tree topology with respect to EX2 and EXEH. 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-$\Gamma$. 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, whereas 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 nonstandard and mixture models are on the opposite side of LG-$\Gamma$, known to be a poor model, whereas 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 with 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 $\Gamma$4 option. EX2, UL3, and EXEO (all three with the $\Gamma$4 option), respectively, use twice, three, and six times as much memory because they are two-level mixtures with two, three, and six matrices, respectively. For example, LG4X and UL3+$\Gamma$4 require 2 GB and 6 GB of memory space, respectively, for the largest TreeBase alignment with 62 taxa and 11,544 sites (accession number M4680). EXEH requires almost 12 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, EXEH 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, whereas EX2, UL3, LG4X, and EXEH, respectively, have two, three, six, and six parameters, thus again requiring additional computing time compared with standard models and LG4M.

Using PhyML-4X with standard options ($\Gamma$4, SPR-based tree searching) on a powerful CPU (Intel® Xeon® E5440 at 2.83 GHz with 16 GB memory) to infer phylogenies for all 84 TreeBase alignments requires about 55 h with standard models (LG was used here), 60 h with LG4M, and 85 h 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 h for EX2, UL3, and EXEH, respectively (total time for EXEH is about 1 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 h for LG, LG4M, LG4X, EX2, UL3, and EXEH, 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; EXEH requires important computing resources for large data sets not only in terms of computing times but also with respect to memory space. The $\Gamma$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 EXEH. However, EXEH (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 table 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 × four-rate models (EX2 + Γ4) and comparable to three-matrix × four-rate models (UL3 + Γ4). However, we do not yet know the change 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.

Acknowledgments

Thanks to Associate Editor Jeffrey 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


