

## **Phylogenetic Mixture Models for Proteins**

Si Quang LE, Nicolas LARTILLOT and Olivier GASCUEL\*

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>

Emails: [le@lirmm.fr](mailto:le@lirmm.fr), [nicolas.lartillot@lirmm.fr](mailto:nicolas.lartillot@lirmm.fr), [gascuel@lirmm.fr](mailto:gascuel@lirmm.fr)

\* Corresponding author

**Phil. Trans. of the Royal Society B, 363:3965–3976, 2008**

## Abstract

Standard protein substitution models use a single amino-acid replacement rate matrix which summarizes the biological, chemical and physical properties of amino acids. However, site evolution is highly heterogeneous and depends on many factors: genetic code, solvent exposure, secondary and tertiary structure, protein function, etc. These impact the substitution pattern, and in most cases a single replacement matrix is not enough to represent all the complexity of the evolutionary processes. This paper explores in a maximum-likelihood framework phylogenetic mixture models, which combine several amino-acid replacement matrices to better fit protein evolution. We learn these mixture models from a large alignment database extracted from HSSP, and test the performance using independent alignments from TreeBase. We compare unsupervised learning approaches, where the site categories are unknown, to supervised ones, where in estimations we use the known category of each site, based on its exposure or its secondary structure. All our models are combined with gamma distributed rates across sites. Results show that highly significant likelihood gains are obtained when using mixture models, compared to the best available single replacement matrices. Mixtures of matrices also improve over mixtures of profiles in the manner of the CAT model. The unsupervised approach tends to be better than the supervised one, but it appears difficult to implement and highly sensitive to the starting values of the parameters, meaning that the supervised approach is still of interest for initialization and model comparison. Using an unsupervised model involving 3 matrices, the average AIC gain per site with TreeBase test alignments is 0.31, 0.49 and 0.61, compared to LG, WAG and JTT, respectively. This 3-matrix model is significantly better than LG for 34 alignments (among 57), and significantly worse for 1 alignment only. Moreover, tree topologies inferred with our mixture models frequently differ from those obtained with single matrices, indicating that using these mixtures impacts not only the likelihood value but also the output tree. All our models and a PhyML implementation are available from <http://atgc.lirmm.fr/mixtures>.

**Keywords:** amino-acid replacement matrices; JTT, WAG and LG; CAT profile model; maximum-likelihood estimations; phylogenetic inference

## Introduction

Amino-acid replacement models are essential in most methods to infer protein phylogenies. In distance methods they are used to estimate the evolutionary distance (*i.e.* the expected number of substitutions per site) between all sequence pairs. In maximum-likelihood and Bayesian methods they are used to compute probabilities of change along the tree branches, and thus the likelihood of the data (see textbooks, *e.g.* Felsenstein 2003, Yang 2006). Standard models use a single amino-acid replacement matrix which summarizes the biological, chemical and physical properties of amino acids. Such  $20 \times 20$  matrix contains estimates of the instantaneous substitution rates from any amino acid to another one. For example, replacements between arginine (positively charged) and aspartate (negatively charged) are under strong negative selection and have low rate, while replacements between isoleucine and valine (both hydrophobic, aliphatic and very non-reactive) are frequent and have high rate (see textbooks, *e.g.* Betts and Russell 2003).

A number of replacement matrices have been proposed since the seminal work of Dayhoff et al. (1972), notably JTT (Jones et al. 1992) and WAG (Whelan and Goldman 2001). Several studies showed that specific matrices should be used for certain analyses, *e.g.* with membrane (Jones et al. 1994) or mitochondrial (Yang et al. 1998) proteins. However, general matrices are usually robust and tend to perform well in many cases, as shown by Keane et al. (2006) for WAG (and to some extent for JTT). Recently, we proposed a new general matrix called LG (after the authors, Le and Gascuel 2008), which significantly improves over previous general matrices. LG was learned from a very large alignment database extracted from Pfam (Bateman et al. 2002), using a maximum-likelihood estimation method that refines Whelan and Goldman's (2001) by incorporating the variability of evolutionary rates across sites in the matrix estimation.

Site evolution is highly heterogeneous and depends on many factors: genetic code, solvent exposure, secondary and tertiary structure, protein function, etc. All these factors impose different constraints on the sites. Some sites are highly conserved, while some others are subject to little pressure and evolve rapidly. This variability of evolutionary rates among sites is well modelled by the use of discrete gamma rate categories (Yang 1993). However, site heterogeneity not only impacts the evolutionary rate but also the substitution pattern. It was shown by several authors (*e.g.* Koshi and Goldstein 1995, Thorne et al. 1996, Goldman et al. 1998) that the substitution pattern differs depending on solvent exposure and secondary structure. Moreover, several models were proposed (*e.g.* Bruno 1996, Koshi and Goldstein 1998, Lartillot and Philippe 2004, Crooks and Brenner 2005)

to account for the fact that depending on their position and role in protein structure and function, sites generally accept only a specific subset of the 20 amino acids. These approaches use sets of models in which the equilibrium frequencies of the 20 amino acids are site-specific. These models rely on simple multinomial processes over the 20 amino-acids, analogous to the F81 (Felsenstein 1981) model of DNA substitution, and are entirely characterised by their amino-acid equilibrium distribution or “profile”. Recently, we learned a series of profile sets with various sizes from a large alignment database extracted from HSSP (Schneider et al. 1997). Our results showed that this empirical profile approach (called CAT, following Lartillot and Philippe 2004) tends to outperform standard replacement matrices, at least with alignments showing a high level of saturation (Le et al. 2008). However, F81-like models are relatively poor as they assume uniform probabilities for mutation from one amino acid to another one, and thus miss a part of the biological constraints acting on site evolution. Better results are expected from models where the mutational processes of each site are modelled using refined (typically general time-reversible) replacement matrices. The purpose of this paper is to study such matrix-based site-dependent models.

Estimating one model per site is not statistically feasible (this would involve too many parameters), and therefore most site-dependent approaches use mixture models. As the most appropriate replacement matrix (or profile) for each site is usually unknown, the likelihood at each site is a weighted average over all alternative matrices (see textbooks, *e.g.* Pagel and Meade 2005, Gascuel and Guindon 2007). When the site category is known (*e.g.* secondary structure or solvent exposure) we can use a partitioning approach which analyses each site with the appropriate matrix. Moreover, several authors proposed to refine the mixture approach using hidden Markov models (*e.g.* Thorne et al. 1996, Felsenstein and Churchill 1996, Goldman et al. 1998) that account for the dependence of site categories (*e.g.* secondary structure) along the sequence. Another refinement of mixtures was proposed by Holmes and Rubin (2002), where the site category can change along the course of time in a way similar to the covarion-like model of Tuffley and Steel (1999).

In this paper we further explore the use of phylogenetic mixture models for proteins in a maximum-likelihood (ML) framework, using up-to-date ML matrix estimation procedures and a very large alignment database, among those that are currently available. In the continuation of (Koshi and Goldstein 1995, Thorne et al. 1996, Goldman et al. 1998) we learn different matrices for the different structural states of the sites (exposed, buried, alpha-helix, beta-sheet and coil). We also estimate matrix mixtures in an unsupervised way, *i.e.* without *a priori* definition of site categories, in a way

close to that of Holmes and Rubin (2002). Contrary to these previous works, all our models incorporate a gamma distribution of site rates. This rate distribution is used to infer trees, as is now usual, but also in the matrix estimations as described in (Le and Gascuel 2008). Results of these new matrix mixtures are compared to those of JTT, WAG and our recent LG and CAT models, using test alignments from TreeBase (Sanderson et al. 1994). In the following, we first describe our data, then the various mixture models and their estimation procedures, and finally compare all these approaches with test alignments.

## **Alignment data sets**

To estimate our mixture models we used HSSP (Homology-derived StructureS of Proteins; Schneider et al. 1997). This database comprises ~35,000 alignments of protein families, each usually containing numerous members (~550 on average). Each alignment is obtained by aligning a protein with known 3-D structure in the Protein Data Bank (PDB), to all its likely sequence homologs in SWISS-PROT. The protein with known structure is named the “test protein” of the alignment; its secondary structure and accessibility to solvent are calculated using DSSP (Kabsch and Sander 1983) and assumed to be representative of the structure of all homologs in the alignment.

HSSP is highly redundant. Typically, a protein may be the test protein of a given alignment and belongs to all alignments corresponding to its homologs with known structure. Moreover, HSSP alignments often contain a huge number of gaps, mainly due to absent or unsequenced domains for some proteins. We thus performed an intensive cleaning of HSSP to extract independent alignments and, within each of the alignments, to select sequences and sites corresponding to well aligned, non-gapped regions. Moreover, we only selected globular proteins, thus discarding membrane proteins that show clearly different patterns of amino-acid replacement (Jones et al. 1994). To eliminate redundancy we used the SWISS-PROT identifiers of proteins; selected alignments do not share any common identifier and correspond to clearly distinct protein families. For each of the retained alignments we selected sequences and sites to obtain a sub-alignment based on a several criteria: presence of the test protein in the sequence set; large number of sequences ( $\geq 10$ ) and sites ( $\geq 100$ ); informative percentage of identities between any sequence pair ([40%, 99%] range); low number of gaps, using GBLOCKS (Castresana 2000) with default options to achieve a final cleaning.

We obtained 1,771 non-redundant sub-alignments (alignments for short from now), with an average of ~56 sequences and ~254 sites per alignment, ~27 million amino-acids in total and very

few gaps (<0.1%). 1,471 alignments were randomly selected for training, while we used the remaining 300 to compare our various models. Using HSSP annotations each site is classified as extended (E), alpha-helix (H), or other (S, T, B, G, I, “.” or “?”). We also classified the sites based on their relative accessibility to solvent (Shrake and Rupley 1973). We used the same two-category partition as in (Goldman et al. 1998) and several other studies, with accessibility threshold equal to 10% and nearly equally weighted buried and exposed categories (~46% and ~54%, respectively). We also used a 3-category partition: the buried class (accessibility < 8%) contains ~40% of the sites, the highly exposed class (accessibility > 45%) contains ~20% of the sites, and the intermediate class ~40% of the sites. This 3-category partition focuses on the highly exposed sites, which are often saturated and appear to have a strong impact on the likelihood value. Additional criteria and details of the selection procedure are described in (Le et al. 2008), and the database is available on request.

To assess the performance of our models we used test alignments from TreeBase (Sanderson et al. 1994). TreeBase contains alignments that have been produced especially for phylogenetic analyses, and thus provide a good benchmark for comparing models meant for phylogenetic reconstruction. Moreover, use of test alignments from a different database should avoid possible biases induced by some feature specific to our HSSP training alignments. Most of TreeBase alignments are carefully aligned with rigorously selected taxa and sequences. These alignments are quite diverse: some are highly cleaned and do not contain any gaps, while some others contain up to 95% gapped sites; some alignments are relatively large, while some others are limited (minimum of 7 and 55 taxa and sites, respectively). All protein alignments from TreeBase (May 2007) were selected, except 3 of them because the set of taxa differed in the alignment and in the published tree, and 2 of them because the maximum pairwise divergence seemed excessively large in a phylogenetic inference context (>2.0 substitutions per site, using a standard WAG distance). Moreover, we removed 5 redundant alignments and 2 very large genomic ones (for computational reasons, with the CAT model). We thus obtained 57 test alignments that should be representative of usual phylogenetic studies, with an average of ~25 sequences and ~550 sites per alignment. These alignments were also used to test our LG replacement matrix (Le and Gascuel 2008) and are available from LG web site: <http://atgc.lirmm.fr/LG> (but removing the 2 very large genomic alignments).

## Mixture models, notation and background

All matrices that we shall discuss comply with the general time-reversible (GTR) model (see textbooks, *e.g.* Felsenstein 2003, Bryant et al. 2005, Yang 2006). Such a matrix contains estimates of the instantaneous substitution rates from any amino acid to another one, and is denoted as  $\mathbf{Q} = (q_{xy})$ , where  $q_{xy}$  is the rate of replacement from  $x$  to  $y$  ( $x \neq y$ ).  $\mathbf{Q}$  can be decomposed into three independent components using

$$\begin{aligned} q_{xy} &= \rho \pi_y r_{x \leftrightarrow y}, \quad x \neq y, \\ q_{xx} &= -\sum_{y \neq x} q_{xy}, \end{aligned} \quad (1)$$

where:  $\rho$  is the global rate of  $\mathbf{Q}$ , equal to the expected number of substitutions per time unit;  $\mathbf{\Pi} = (\pi_x)$  is the vector of amino-acid equilibrium frequencies;  $\mathbf{R} = (r_{x \leftrightarrow y})$  is the (symmetric) exchangeability matrix, which represents the general propensity of exchanges between amino-acids, independently of the amino-acid frequencies within the studied sequences (represented by  $\mathbf{\Pi}$ ). In the following, we assume that  $\mathbf{R}$  is normalized (*i.e.*  $-\sum \pi_x q_{xx} = \rho$ ), and thus  $\mathbf{Q}$  contains  $1(\rho) + 19(\mathbf{\Pi}) + 189(\mathbf{R}) = 209$  free parameters to be estimated from the data. When a single replacement matrix is used (*e.g.* WAG), it is normalized (*i.e.*  $\rho = 1$ ) to obtain a simple branch length interpretation in terms of number of expected substitutions per site. When several matrices are used, they are no longer normalized (*e.g.* exposed sites evolve about thrice as fast as buried sites) and the model requires a specific global normalisation that is discussed below.

Amino-acid changes over the course of time are represented by the matrix  $\mathbf{P}(t) = (p_{xy}(t))$ , where  $p_{xy}(t)$  is the probability of observing a change from  $x$  to  $y$  when the elapsed time is  $t$ . The probability  $p_{xy}(dt)$  of changing from  $x$  to  $y$  ( $x \neq y$ ) in infinitesimal time  $dt$  is equal to  $q_{xy}dt$ . This implies the following basic relationship between the substitution rates and probabilities of change:

$$\mathbf{P}(t) = e^{\mathbf{Q}t}, \quad (2)$$

where the right term denotes the matrix exponential.

Assuming a single replacement matrix  $\mathbf{Q}$  and no variability of rates among sites, the likelihood of the data (denoted  $D$ ) for a given tree  $T$  (including branch lengths) is:

$$L(T, \mathbf{Q}; D) = \prod_i L(T, \mathbf{Q}; D_i), \quad (3)$$

where the product runs over all the sites (independence assumption), and where  $L(T, \mathbf{Q}; D_i)$  is the likelihood of the data at site  $i$  (denoted  $D_i$ ) given  $T$  and  $\mathbf{Q}$ .  $L(T, \mathbf{Q}; D_i)$  is computed by applying Equation (2) to each tree branch ( $t$  is the branch length) and using the pruning algorithm (Felsenstein 1981).

However, it is acknowledged that sites do not evolve at the same rate due to various evolutionary pressures. In the ML framework, practical implementations rely on a simple mixture model with discrete categories of rates. Each site belongs to a category  $c \in \{1, 2, \dots, C\}$  with rate  $\rho_c$ . Yang's (1993) approach involves categories with identical probabilities (equal to  $1/C$ ) and  $\rho_c$  rates being defined by the parameter  $\alpha$  of a gamma distribution, which is usually fitted to the analysed data set. The likelihood of the data for tree  $T$ , replacement matrix  $\mathbf{Q}$  and gamma distributed rate categories is:

$$L(T, \mathbf{Q}, \alpha; D) = \prod_i \sum_{1 \leq c \leq C} \frac{1}{C} L(T, \rho_c \mathbf{Q}; D_i). \quad (4)$$

In the mixture defined by Equation (4), we have  $C$  replacement matrices that only differ by their global rates. In this paper we consider the more general setting where site categories (*e.g.* buried/exposed) correspond to different substitution patterns, each modelled using a different replacement matrix. Let  $\Theta$  denote the set of substitution pattern categories,  $\theta$  be a pattern category of  $\Theta$ ,  $\mathbf{Q}_\theta$  the replacement matrix corresponding to  $\theta$ ,  $\pi_\theta$  the *a priori* probability of  $\theta$ , and  $\rho_\theta$  the global rate of  $\mathbf{Q}_\theta$ . In the partition approach, the category of each site  $i$  is known and each site is analysed with the proper replacement matrix (*e.g.* see Gascuel and Guindon 2007). However, site categories are not always known, or may be known with a large uncertainty (*e.g.* secondary-structure states are somewhat arbitrary and non-fully conserved among homologous proteins). Moreover, we will discuss models where site categories do not have any obvious interpretation and are learned empirically from the data (the unsupervised way). Mixture models are used to cope with such cases. For each site we sum over all possible categories, and the likelihood of the data is expressed as:

$$L(T, \Theta, \alpha; D) = \prod_i \left[ \sum_{1 \leq \theta \leq |\Theta|} \pi_\theta \sum_{1 \leq c \leq C} \frac{1}{C} L(T, \rho_c \mathbf{Q}_\theta; D_i) \right]. \quad (5)$$

Equation (5) defines a mixture with  $|\Theta|$  (number of patterns)  $\times$   $C$  (number of rates) categories, each with probability  $\pi_\theta/C$  and replacement matrix  $\rho_c \mathbf{Q}_\theta$ . All mixture models discussed in this paper comply with Equation (5). The differences come from the number of pattern categories, the properties

of the  $\mathbf{Q}_\theta$  matrices (GTR or F81), and the way these models were learned from the data. The gamma distribution of rates is assumed to be the same among pattern categories, as we did not observe any significant improvement when using several rate distributions.

In all models (except CAT, see below) the proportions  $\pi_\theta$  of pattern categories are fitted to the analysed data set. This is an important feature as proteins are highly heterogeneous. For example, some proteins contain alpha-helices and no beta-sheets, while others contain beta-sheets only, and finally some contain both. Analysing all proteins with fixed proportions of alpha-helices and beta-sheets would poorly fit this biological reality and result in a loss of likelihood value. Mixtures defined by equation (5) thus require  $|\Theta|-1$  additional free parameters, compared to Yang's model in equation (4). Moreover, we have to normalize the mixture to ensure branch-length interpretation. In practice, since the  $\pi_\theta$  proportions vary from one data set to another one, we do not normalize the mixture but rescale the inferred tree, which is equivalent. This post-processing involves multiplying every branch length by the expected rate of the mixture (*i.e.*  $\sum \pi_\theta \rho_\theta$ ), after all parameters ( $\pi_\theta$ , topology, branch-lengths, etc.) have been estimated from the data.

The CAT model (Lartillot and Philippe 2004; Le et al. 2008) uses simplified F81-like replacement matrices (all exchangeabilities  $r_{x \leftrightarrow y}$ ,  $x \neq y$ , are equal). Each matrix is thus defined by an amino-acid profile (19 free parameters) corresponding to the equilibrium distribution of the substitution process, and a number of profiles (up to 60 in our experiments) are used to accurately model amino-acid substitutions. This model is intended to fit the common observation that sites only contain a few amino-acids corresponding to precise biochemical constraints, even in case of saturation. Thus, the CAT profiles contain a few amino-acids with significant probability, while the other amino-acids have nearly-zero probability. In (Le et al. 2008) it was not clear whether adjusting the profile proportions (the  $\pi_\theta$  parameters) results in an increase in likelihood sufficient to compensate for the high number of additional parameters (59 with 60 profiles). We thus preferred to use fixed profile proportions, and the same holds in this paper.

In tree inference, all models simply use Equation (5) to compute the tree likelihood and optimize the topology, branch lengths and model parameters. The computational cost depends greatly on the number of mixture categories. In standard implementations, both the memory requirement and the computing time are nearly proportional to the number of categories (see textbook, *e.g.* Bryant et al. 2005). For example, using a 3-matrix mixture model with 4 gamma categories should be 3 times slower than using a single matrix, and 12 times slower than using a single matrix without rates across

sites. Moreover, the same holds for memory requirement, which may be problematic for large data sets. The problem with CAT (up to  $60 \times 4$  categories) is partly alleviated using implementation refinements (Lartillot and Philippe 2004). Assume that a profile corresponds to the aliphatic amino-acids; to analyse the data with this profile we only need a  $4 \times 4$  matrix corresponding to I, L, V and Other amino-acids. This appreciably reduces the computing time and memory requirement, but in our current implementation CAT is still slow (*e.g.* CAT with 60 profiles is  $\sim 8$  times slower than a 3-matrix mixture). Moreover, mixtures often require more iterations to optimize the parameters than single matrices (*e.g.* running a 3-matrix mixture is  $\sim 4$  times slower than a single matrix as WAG or JTT).

In matrix estimation, we shall see that to accelerate the computation Equation (5) can be simplified without loss of accuracy. Moreover, a specific EM algorithm was designed to learn CAT profiles (see Le et al. 2008).

## **The supervised approach: estimation procedure and models**

The supervised approach involves using available knowledge to guide the learning procedure. Here, we know (*e.g.* Koshi and Goldstein 1995, Goldman et al. 1998) that the replacement process differs depending on secondary structure and solvent exposure. We thus estimated different replacement matrices for several site partitions, based on the information available in HSSP. Three models were learned:

- EX2 is a 2-matrix model corresponding to exposed/buried sites (see above).
- EX3 is a 3-matrix model corresponding to highly-exposed/intermediate/buried sites.
- EHO is a 3-matrix model corresponding to extended/helix/other sites.

The same estimation procedure was used for these three models. This procedure is closely related to that used to estimate our LG matrix (Le and Gascuel 2008) and is summarized below with EX2.

- (1) For every alignment  $D^a$  in the training database, estimate a phylogenetic tree  $T^a$  using PhyML (Guindon and Gascuel 2003) with LG with four gamma categories ( $\Gamma 4$  option).
- (2) For every site  $i$  in  $D^a$ , classify  $i$  into the rate category with maximum *a posteriori* probability (MAP); let  $\rho(i)$  denote the corresponding rate.

- (3) Using HSSP annotations divide the training database into exposed/buried sites and separately estimate a matrix for each category using XRATE (Klosterman et al. 2006) and the inferred  $T^a$  trees. Just as with LG, do not use standard site likelihood (4) summing over all rate categories, but simply use the MAP rate category, that is:

$$L(T^a, \mathbf{Q}, \alpha^a; D_i^a) = L(T^a, \rho(i)\mathbf{Q}; D_i^a). \quad (6)$$

In this equation,  $T^a$  and  $\rho(i)$  are fixed and only the replacement matrix  $\mathbf{Q}$  has to be estimated from the data. All parameters of  $\mathbf{Q}$  ( $\rho$ ,  $\mathbf{\Pi}$  and  $\mathbf{R}$ , see equation (1)) are estimated by ML using XRATE.

- (4) For each site category, compute the expected *a priori* rate by averaging the  $\rho(i)$  values, and multiply the global rate found by XRATE by this average rate. This operation rescales the two matrices so that their global rates are comparable and they can be applied to the same trees. Moreover, to help interpretability, normalize the mixture using the constraint  $\sum \rho_\theta \pi_\theta = 1$ , where  $\theta$  is exposed/buried and  $\pi_\theta$  is the global proportion of exposed/buried sites in the training set.
- (5) Go to (1) and iterate this estimation procedure, but use the exposed and buried matrices and a site partition in place of LG. Steps (2), (3) and (4) remain identical, and the procedure is repeated until convergence. However, the  $T^a$  trees are inferred only once using the partition model (during the second iteration), since this part of the computation is very heavy. Further iterations use the nearly optimal trees thus obtained, which are sufficient to obtain accurate matrix estimates (Whelan and Goldman 2001, Le and Gascuel 2008). Three iterations were enough for all models.

XRATE is able to deal with the standard mixture equation (4), instead of MAP equation (6), at least when a unique discrete rate distribution is chosen for all training alignments. But we observed that using (6) is much faster, less affected by local optima and tends to provide better results (Le and Gascuel 2008). This is why we adopted the same strategy here, which is close to Viterbi's approximation that provides very good results when estimating HMMs (Durbin et al. 1998). Running times of this supervised scheme are relatively high, due to the size of the training data set, but acceptable; *e.g.* ~2 days of computation were needed to estimate the EX3 model using our cluster (16 X 2.33GHz biprocessors with 8 Gb RAM).

The main features of the matrices thus estimated are summarized in Table 1. We see that:

- The global rate is quite different for the exposed and buried categories: with EX3 the buried category is about twice as slow as the intermediate category, while the highly-exposed category is almost twice faster. This contrast is higher than that observed by Goldman et al. (1998) using a counting estimation procedure. They found a ratio of about 2.08 between exposed and buried rates (EX2 site partition), compared to 2.45 here, probably due to the fact that counting tends to underestimate the number of hidden substitutions compared with ML estimations. We also find a slightly higher contrast than found by Goldman et al. with secondary-structure categories (*e.g.* 1.36 versus 1.28 for Helix/Extended), but our results confirm their main conclusion that the global substitution rate does not change much among secondary-structure categories.
- The correlations between the amino-acid frequencies and exchangeabilities of the various matrices and those from LG indicate clear differences in the substitution patterns (see also matrices and graphics on <http://atgc.lirmm.fr/mixtures>). Mainly, we see that the amino-acid equilibrium frequencies are quite different among site categories. For example, as expected the buried category mostly contains hydrophobic amino-acids, while the helix category contains a large proportion of alanines but very few glycines. The correlations between exchangeabilities are much higher and above 0.9, except with highly exposed sites, which represent a clearly distinct site category. This high correlation level is explained by the fact that exchangeabilities represent the general propensities of amino acids, which are relatively invariant among categories (notably secondary-structure ones), while amino-acid frequencies largely account for the local constraints acting on sites. However, we shall see that the impact on tree likelihood of these moderate differences in exchangeabilities across categories is similar to that of the large differences in amino-acid frequencies. Indeed, we see from Table 1 that the correlations between the whole Q matrices mostly depend on exchangeabilities rather than on amino-acid frequencies.

### **The unsupervised approach: estimation procedure and models**

In the unsupervised approach, we ignore available knowledge and try to directly infer new site partitions from the data, along with the corresponding replacement matrices. In principle, this should lead to better models than the supervised approach, as we have more degrees of freedom and can still recover the known site partitions. But the unsupervised approach involves complex numerical optimization with a number of local optima. An intermediate way is thus to use a semi-supervised procedure, where the starting solution is obtained using a known site partition and the procedure

described above, and then to refine this model in an unsupervised way. We performed several experiments along these lines to estimate two-category and three-category models, which we call UL2 and UL3 (Unsupervised Learning), respectively. We implemented two basic estimation strategies, both close to above supervised scheme. The first strategy (called mixed-strategy) uses the ability of XRATE to deal with mixtures, in combination with our Viterbi-like approximation (6). It uses the same 5 steps as the supervised procedure, and we only detail the differences:

- (1) Infer the  $T^a$  trees using PhyML, as in the supervised procedure.
- (2) Classify every site in the MAP rate category, as in the supervised procedure.
- (3) Use an appropriate phylogrammar in XRATE to define a mixture of  $|\Theta|$  matrices ( $\mathbf{Q}_\theta$ ) with proportions ( $\pi_\theta$ ) estimated from the data but identical for all training alignments (XRATE does not allow for different category proportions among alignments). Then, run XRATE with site likelihood:

$$L(T^a, \Theta, \alpha^a; D_i^a) = \sum_{\theta} \pi_{\theta} L(T^a, \rho(i) \mathbf{Q}_{\theta}; D_i^a), \quad (7)$$

which is similar to (6) in that we do not sum over rate categories but use the MAP rate. However, we sum over pattern categories, and the estimation procedure is slow, even with UL2 (~2 days of computation on our cluster, for each XRATE run).

- (4) Normalize the matrices using the constraint  $\sum_{\theta} \pi_{\theta} \rho_{\theta} = 1$ .
- (5) Go to (1) and iterate the learning procedure until convergence (three iterations were enough in all experiments), but use the estimated mixture to infer the trees in place of LG and initialize XRATE in step (3). Other steps remain identical.

For the first iteration, we need starting matrices in step (3). Several starting points were used in our experiments. For UL2, we started from EX2 matrices and from 2 matrices with uniformly randomly drawn exchangeabilities and amino-acid frequencies; best results were obtained with EX2. For UL3, we started from EX3 and 3 random matrices; best results were obtained with the random matrices. The fit of the various solutions was compared with TreeBase test alignments (Table 2).

The second strategy (called MAP-strategy) is even closer to the supervised scheme and does not use any mixture within XRATE. However, it requires starting from an initial mixture of  $\mathbf{Q}_\theta$  matrices.

- (1) Infer the  $T^a$  trees using PhyML with  $\Gamma 4$  option and the mixture of  $\mathbf{Q}_\theta$  matrices; the  $\pi_\theta$  proportions are optimized for each alignment separately (in this respect, the MAP-strategy is more flexible than the mixed-strategy).
- (2) For every site  $i$ , classify  $i$  in the MAP rate and pattern categories.
- (3) One matrix is learned for each pattern category separately, as in the supervised procedure.
- (4) Normalize the matrices, as in the supervised procedure.
- (5) Go to (1) and iterate until convergence (three iterations were enough in all experiments).

We tested two starting mixtures. For UL2 (UL3), we used EX2 (EX3) and the 2-category (3-category) mixture obtained by the mixed-strategy. In both cases, best results were obtained by combining the two strategies. The mixed-strategy provided an initial mixture, which was significantly improved using the MAP-strategy, likely due its greater flexibility. Results with TreeBase are displayed in Table 2, and the same method ordering was obtained with HSSP test alignments (not shown), meaning that our model choice is not biased in favour of TreeBase. We see large differences between the two strategies and the various starting points. This suggests that other combinations could be tested, which would likely improve our current models. However, running any of these approaches requires important computational resources; *e.g.* about 12 days of computation on our cluster were needed to estimate the UL3 model with the mixed-strategy and a random starting point.

The main features of our best UL2 and UL3 models (obtained by combining the mixed- and MAP-strategies) are displayed in Table 3 (see also <http://atgc.lirmm.fr/mixtures>). UL2 matrices are denoted as M1 and M2, while UL3 ones are denoted Q1, Q2 and Q3. We see that:

- Unsupervised models are more distant from LG than supervised ones; *e.g.* all exchangeability correlations are below 0.9, while all supervised ones (but one) are above 0.9. LG represents the average model, and thus unsupervised models are more varied than supervised ones. This can be explained by two factors: (1) the supervised categories are somewhat imprecise and not fully reliable, meaning that they contain “average” sites that could be classified in other categories as well; (2) the unsupervised scheme tends to exacerbate the differences between the mixture categories.
- UL2 is strongly correlated with EX2: M1 is quite close to the exposed matrix (proportion, global rate, frequencies and exchangeabilities), and the same holds with M2 and the buried matrix. UL2 is obtained starting from EX2 and combining the two estimation strategies, but this high

correlation still holds when starting from random matrices (not shown), and similar results have been found by Holmes and Rubin (2002) using a different unsupervised approach. This means that the main factor in amino-acid substitution is accessibility to solvent, which corresponds to a known, well-documented fact. This affects the global substitution rate ( $\rho_\theta$ ), but also the substitution process ( $\mathbf{R}_\theta$ ) and (obviously) the amino-acid equilibrium frequencies ( $\mathbf{\Pi}_\theta$ ), which tend to be hydrophobic/hydrophilic depending on site exposure. As expected, M1 is close to the “other” secondary-structure category (containing the turns and coils, which are typically exposed), while M2 is close to “extended” sites (the most buried secondary-structure category).

- UL3 is more difficult to interpret than UL2. Q1 is relatively close to the exposed (or even highly exposed) matrix, Q3 is relatively close to the buried matrix, while Q2 does not correlate with exposure-based matrices but is (relatively) close to the “other” secondary-structure matrix. A similar interpretation of Q2 is found with HSSP alignments when looking at the true structural categories of the sites; when Q2 is the mixture category with highest posterior probability, about half of the sites are buried and half exposed, while ~60% of the sites are in the “other” secondary-structure category (versus ~43% in average). Most notably, Q2 amino-acid frequencies show a very high proportion of glycines and prolines (~30%, versus ~10% in average), which are unique amino acids in that they influence the conformation of the polypeptide and are often found in turns. Moreover, Q2 is well conserved with a low global rate, though it is closer to the exposed model than to the buried one. This site category was not found by Holmes and Rubin (2002) when testing their model with 3 (and 4) categories; they observed a third “tiny” category favouring alanine, glycine and serine (all of which have very small side-chains), in addition to the exposed and buried categories. Thus, UL3 seems to combine exposure and secondary-structure information (in an efficient way, as we shall see in the Result section). Further investigations would deserve to be conducted to better understand the evolutionary and biochemical properties of UL3 site categories and replacement matrices.

## Results with test alignments

We used the 57 TreeBase test alignments to compare supervised (EHO, EX2 and EX3) and unsupervised (UL2 and UL3) matrix mixture models, to single matrices (JTT, WAG and LG) and profile mixtures (CAT20 and CAT60 with 20 and 60 profiles, respectively; see Le et al. (2008) for details). Note that the supervised models, which were learned using HSSP-based site partitions, are

used here as mixtures, as we do not have structural information in TreeBase (and in number of phylogenetic data sets). All models were run with PhyML using 4 gamma rate categories ( $\Gamma_4$ ), BioNJ (Gascuel 1997) starting tree and SPR-based tree topology search (Hordijk and Gascuel 2005). This imposed some adaptations of standard PhyML. In the current implementation, an initial ML tree is first inferred with LG+ $\Gamma_4$  in the usual way, and then the mixture is used to refine this first tree, with the model parameters (mixture proportions and gamma shape parameter) being adjusted along the way.

For all models, we measured the AIC criterion (Akaike 1974) on each of the test alignments:

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

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 parameters of model M. Single matrix and CAT models involve the same number of parameters (number of branches plus one, corresponding to the gamma shape parameter), UL2 and EX2 require one additional parameter (mixture proportion), while EHO, EX3 and UL3 require two additional parameters (two mixture proportions). We computed the average AIC per site of model M for all test alignments, which is simply

$$AIC/site(M) = \sum_a AIC(M, D^a) / \sum_a s^a, \quad (8)$$

where  $s^a$  is the number of sites in  $D^a$ . All models were compared to LG using criterion (8). To complete this global average result, we also counted the number of alignments where  $AIC(M, D^a)$  is better/worse than  $AIC(LG, D^a)$ . Moreover, to assess the statistical significance of the observed difference between M and LG, we used a Kishino-Hasegawa (KH; 1989) test with  $p < 0.01$ .

For each of the inferred trees, we measured the tree length (sum of branch lengths) and the gamma shape parameter, as best models tend to produce longer trees capturing more hidden substitutions (see Pagel and Meade, 2005, for a discussion on tree length and likelihood value). We also compared the topology of inferred trees. The true tree is not known with real data (as opposed to simulated data), and our aim was to measure the impact of the various models in terms of topology, *i.e.* whether we frequently infer a different tree topology when improving the substitution model. Indeed, it is commonly believed that tree topologies inferred with usual models (JTT, WAG, etc) tend to be identical, which would mean that any efforts to refine these models are somewhat useless. When different topologies are found, we should prefer the one with best likelihood value. However, the

difference may be slight and non-significant, so we cannot reject the topology with the lower likelihood value. Thus, we counted the number of alignments where the tree built using any given model M is not the same as the tree inferred with LG, and the significance of these topological differences was assessed using a KH test ( $p < 0.01$ ).

Average AIC results are displayed in Figure 1, and Figure 2 provides the number of alignments where each model is (significantly) better/worse than LG. We see that:

- LG clearly outperforms JTT and WAG (as shown in Le and Gascuel 2008), but is outperformed by the mixture models. While LG is often significantly better (and rarely worse) than JTT and WAG, we observe the converse with mixtures (but CAT), which are often significantly better than LG and rarely worse; *e.g.* compared to UL3 (the best mixture), LG is significantly better with 1 alignment only, while UL3 is significantly better than LG with 34 alignments (among 57). Moreover, the AIC gain of matrix mixtures is largely due to the different exchangeabilities among sites categories, and not only to the differences in amino-acid composition. For example, we ran a 2-category mixture with EX2 frequencies (highly contrasted between exposed and buried sites), but LG exchangeabilities for both categories; the AIC gain of this model is 0.074, compared to 0.151 with the full EX2 model.
- CAT results are a bit disappointing, as the average performance of CAT20 is only slightly better than that of LG, and as both CAT20 and CAT60 are often worse than LG, and even significantly worse (with 10 and 5 alignments, respectively). However, CAT was designed for saturated data sets, and we showed that it performs well with such data (Le et al. 2008). To this purpose, we used the saturation index defined by Lartillot et al. (2007), which corresponds to the parsimony-based number of convergences and reversions. When looking (Figure 1) at the alignments with saturation index per site larger than 2, we see that CAT20 has nearly the same performance as the 2-matrix models (EX2 and UL2). This makes sense as all these models involve about the same number of numerical values (~400 rates and probabilities) and thus similar amount of knowledge. Moreover, CAT60 is nearly as good as UL3 (1199 and 630 numerical values, respectively). Above all, we see that with such saturated data the contrast between all models is much increased; *e.g.* the difference between UL3 (best model) and JTT (worse model) is ~1.7 AIC point per site, meaning that with 300 sites the difference is as large as ~500 AIC points, which is considerable.

- Supervised models show similar performance. When all alignments are considered, exposure-based models (EX2 and EX3) slightly improves secondary-structure-based model (EHO), while EX2 and EX3 are close. Thus, EX2 should be preferred for most studies as it requires less computing time and memory.
- UL2 is also quite close to EX2, as expected since both are strongly correlated. As EX2 is based on known properties of proteins, and thus is easily interpretable, we believe that it should be preferred over UL2 in most cases. However, the unsupervised approach demonstrates its advantage with 3-matrix models, since UL3 clearly outperforms EHO and EX3. It also improves CAT60, despite the fact that it contains fewer numerical values, run faster (~8 times) and requires less memory. Unsupervised mixtures of matrices thus seem to be an efficient and accurate way to encode the main features of amino-acid replacements.

Table 4 provides the main features of the trees inferred using the various models. We see that:

- Though LG trees are longer than JTT and WAG trees, mixture trees are even longer. This is a clear tendency for all mixtures, indicating that these models tend to infer more hidden substitutions than LG (and JTT and WAG). The gamma shape parameter has a different behaviour. The variability of rates among sites tends to be lower ( $\alpha$  is higher) with WAG than with LG (JTT and LG are nearly identical), and the converse is observed with CAT models, which is consistent with our observations with tree length, as evolutionary distances and branch lengths are increased when the  $\alpha$  value decreases. With matrix mixtures the  $\alpha$  value is higher than with LG, but this cannot be interpreted as a lower variability of rates, because part of this variability is accounted for by the variable global rates of the matrices defining the mixtures.
- All models often infer topologies that differ from LG topology, and the topological differences frequently correspond to significant likelihood gains (mixture models), or losses (JTT and WAG), compared to LG. Moreover, the topological distance between trees is also high, especially with CAT. In fact, CAT often (~85% of alignments) infers trees that clearly (~25% of clades) differ from LG trees. The exact reasons for these marked differences between CAT and LG remain to be investigated. A possibility is that it corresponds to more accurate inferences. In this direction, we showed (Le et al. 2008) in some well documented case studies that CAT has a good resistance to long-branch attraction artefacts. Alternatively, the use of F81 processes could cause a loss of accuracy of CAT. All together, our results show that protein substitution modelling impacts the topology of inferred trees. Although it is not clear whether the resulting topologies are closer to

the true topologies, they are different from the topologies inferred using standard models (JTT, WAG), with higher likelihood values in most cases, and thus these alternative topologies should be of great interest for phylogeneticists.

## Discussion

Our results with test alignments show that highly significant likelihood gains are obtained using mixture models, compared to single replacement matrices (JTT, WAG and LG). Unsupervised models tend to outperform supervised ones, but only slightly so with 2 matrices, as the main factor in amino-acid replacement seems to be the accessibility to solvent. With 3 matrices, our unsupervised model (UL3) combines exposure and secondary-structure aspects and is clearly the best model we tested; the average AIC gain per site with TreeBase test alignments is 0.31, 0.49 and 0.61, compared to LG, WAG and JTT, respectively. Moreover, UL3 is significantly better than LG for 34 alignments (among 57), and significantly worse with 1 alignment only. CAT performs well with saturated alignments (with 60 profiles, CAT60 is nearly as good as UL3), but is slow in an ML context due to the number of site categories (profiles). However, this limitation is alleviated in the Bayesian framework (thanks to data augmentation) where CAT60 should be quite relevant.

As said in the Introduction, the work presented here is a continuation of previous researches, mainly by Thorne et al. (1996) and Goldman et al. (1998) for the supervised scheme, and Holmes and Rubin (2002) for the unsupervised one. Moreover, Koshi and Goldstein (*e.g.* 1995, 1998) explored similar questions and models using Bayesian approaches. The main differences with these previous works are:

- The size of our training database, which matters to estimate such large number of parameters; *e.g.* we showed in Le and Gascuel (2008) that about half of the gain of LG compared to WAG was induced by the training alignments.
- The use of up-to-date ML programs (XRATE for matrices and PhyML for trees), while Thorne, Goldman and colleagues used NJ (Saitou and Nei 1987) and counting-based matrix estimations (and an EM algorithm to estimate their HMM model), and Holmes and Rubin inferred the phylogenies using NJ. Moreover, we iterated the learning procedure, *i.e.* repeatedly inferred the trees using the mixture and estimated the mixture parameters using these trees. The gain was appreciable with LG (Le and Gascuel 2008), but is higher with mixtures; *e.g.* with UL3, the

mixture obtained after the first iteration of the mixed-strategy has an AIC gain per site with TreeBase of 0.25 compared to LG, while the final UL3 model has 0.31.

- The fact that all our models include a gamma distribution of rates, which is accounted for in model estimation. This feature explains the second half of LG's gain compared to WAG (Le and Gascuel 2008). Moreover, modelling rates across sites in tree inference is of first importance, even with pattern-based site categories with variable global rates (*e.g.* buried/exposed). Preliminary experiments show that our models clearly outperform PASSML (Lio et al. 1998), which implements Thorne's and Goldman's structural models, but does not explicitly incorporate rates across sites as in our equation (5).
- The use of mixtures with category proportions that are adjusted to the analysed alignment. This is a simple approach (simpler than HMM and Markov Modulated Markov models used in some of these previous researches), but it fits important features of proteins which are highly heterogeneous.

All of this (very large database, full ML approach, refined models) has been possible thanks to today computers and the recent algorithmic developments on ML estimation of phylogenetic models (trees and replacement matrices). All together, we obtained simple, robust and ready-to-use models. A PhyML implementation is available from <http://atgc.lirmm.fr/mixtures/>.

Further investigations should include: refinements of these models, *e.g.* using more site categories or non-parametric rate distributions; assessment of these models in a Bayesian framework; comparison with partition models, when the structure of the studied protein is known; algorithmic refinements of the unsupervised estimation procedure, to cope with the multiple local optima that we observed; estimation of mixture models specific to certain protein groups (*e.g.* mitochondrial or membrane proteins) or life domains (*e.g.* viruses or apicomplexa).

## **Acknowledgements**

Sincere thanks to Nick Goldman, Stéphane Guindon, Ian Holmes, Simon Whelan, Ziheng Yang, Avril Coghlan and two anonymous reviewers for their help, suggestions and comments. This work was supported by ANR BIOSYS (MitoSys project).

## References

- Akaike H. 1974. A new look at statistical model identification. *IEEE Transactions on Automatic Control* AU-19:716-722.
- Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer ELL. 2002. The Pfam protein families database. *Nucleic Acids Res.* 30:276-280. <http://pfam.cgb.ki.se/>
- Betts MJ, Russell RB. 2003. Amino acid properties and consequences of substitutions. In: Barnes MR, Gray IC, editors. *Bioinformatics for Geneticists*. Wiley, New York.
- Bruno WJ. 1996. Modeling residue usage in aligned protein sequences via maximum likelihood. *Mol. Biol. Evol.* 13:1368–1374.
- 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.
- Crooks GE, Brenner SE. 2005. An alternative model of amino-acid replacement. *Bioinformatics* 21:975–980.
- 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.
- 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 method of sequence data. *Mol. Biol. Evol.* 14:685–695.

- Gascuel O, Guindon S. 2007. Modelling the variability of evolutionary processes. In O. Gascuel and M. Steels, editors, *Reconstructing Evolution: new mathematical and computational advances*, Oxford University Press, 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, Gascuel O. 2003. A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* 52:696–704.
- Holmes I, Rubin GM. 2002. An expectation maximization algorithm for training hidden substitution models. *J. Mol. Biol.* 317:753–764.
- Hordijk W, Gascuel O. 2005. Improving the efficiency of SPR moves in phylogenetic tree search methods based on maximum likelihood. *Bioinformatics* 21(24):4338-4347.
- Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *CABIOS* 8:275–282.
- Jones DT, Taylor WR, Thornton JM. 1994. A mutation data matrix for transmembrane proteins. *FEBS Lett.* 339:269-275.
- Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577-637.
- Keane TM, Creevey CJ, Pentony MM, Naughton TJ, McInerney 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 PS, Uzilov AV, Bendaña 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 (1):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.

- 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.
- Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evolutionary Biology* 7 Suppl 1:S4.
- Le Si Q, Gascuel O. 2008. An Improved General Amino-Acid Replacement Matrix. *Mol. Biol. Evol.* 25:1307-1320.
- Le Si Q, Gascuel O, Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction, submitted manuscript.
- Lio P, Goldman N, Thorne JL, Jones DT. 1998. PASSML: combining evolutionary inference and protein secondary structure prediction. *Bioinformatics* 14:726-733.
- 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.
- Robinson D, Foulds L. 1979. Comparison of weighted labeled trees. *Lect. Notes Math.* 748:119-126.
- Saitou N, Nei M. 1987. The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees. *Mol. Biol. Evol.* 4:406-425.
- 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(6):183. <http://www.treebase.org/>
- Schneider R, de Daruvar A, Sander C. 1997. The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res.* 25:226-230.
- Shrake A, Rupley JA. 1973. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.* 79:351-372.
- Thorne JL, Goldman N, Jones DT. 1996. Combining Protein Evolution and Secondary Structure. *Mol. Biol. Evol.* 13:666-673.
- Tuffley C, Steel M. 1998. Modeling the covarion hypothesis of nucleotide substitution. *Math. Biosci.* 147:63–91.
- 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 Univ. Press, Oxford, UK.

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

	EX2		EX3			EHO		
$\theta$	Exposed	Buried	HExposed	Intermediate	Buried	Extended	Helix	Other
$\pi_\theta$	0.552	0.448	0.196	0.389	0.415	0.208	0.360	0.432
$\rho_\theta$	1.360	0.557	1.897	1.046	0.534	0.857	1.163	0.932
$\mathbf{\Pi}_\theta/\text{LG}$	0.619	0.673	0.468	0.832	0.659	0.729	0.777	0.625
$\mathbf{R}_\theta/\text{LG}$	0.927	0.934	0.856	0.942	0.925	0.964	0.935	0.958
$\mathbf{Q}_\theta/\text{LG}$	0.940	0.915	0.877	0.955	0.904	0.959	0.946	0.961

**Table 1: Replacement matrices for solvent-exposure and secondary-structure site categories**

Note:  $\pi_\theta$ : proportion of  $\theta$  in the training set;  $\rho_\theta$ : global rate of  $\theta$ ;  $\mathbf{\Pi}_\theta/\text{LG}$ : correlation of  $\theta$  amino-acid frequencies compared to LG frequencies;  $\mathbf{R}_\theta/\text{LG}$ : correlation of  $\theta$  and LG exchangeabilities using log values (exchangeabilities are highly contrasted with some very small values);  $\mathbf{Q}_\theta/\text{LG}$ : correlation of  $\theta$  and LG rates using log values; HExposed: highly exposed sites.

	MIXED EX START	MIXED RANDOM START	MAP EX START	MAP MIXED START
UL2	0.145	0.135	0.156	0.180
UL3	0.250	0.282	0.223	0.306

**Table 2: Comparison of unsupervised learning strategies**

Note: The model fit is measured with 57 TreeBase test alignments, using the AIC per site gain compared to LG (see Results section for details). The higher the gain, the better is the model. MIXED: mixed-strategy; MAP: MAP-strategy; EX START: EX2 or EX3 models are used as starting points; RANDOM START: starting matrices are randomly drawn; MIXED START: use as starting point the best model obtained by the mixed-strategy.

	UL2		UL3		
$\theta$	M1	M2	Q1	Q2	Q3
$\pi_\theta$	0.498	0.502	0.320	0.282	0.398
$\rho_\theta$	1.27	0.735	1.647	0.702	0.690
$\mathbf{\Pi}_\theta/\text{LG}$	0.527	0.677	0.545	0.454	0.688
$\mathbf{R}_\theta/\text{LG}$	0.824	0.891	0.788	0.827	0.836
$\mathbf{\Pi}_\theta/\text{SUP}$	0.984 e 0.854 O	0.969 b 0.842 E	0.935 e 0.886 H	0.273 e 0.709 O	0.926 b 0.873 E
$\mathbf{R}_\theta/\text{SUP}$	0.955 e 0.919 O	0.932 b 0.915 E	0.897 e 0.854 O	0.809 e 0.837 O	0.852 b 0.845 E

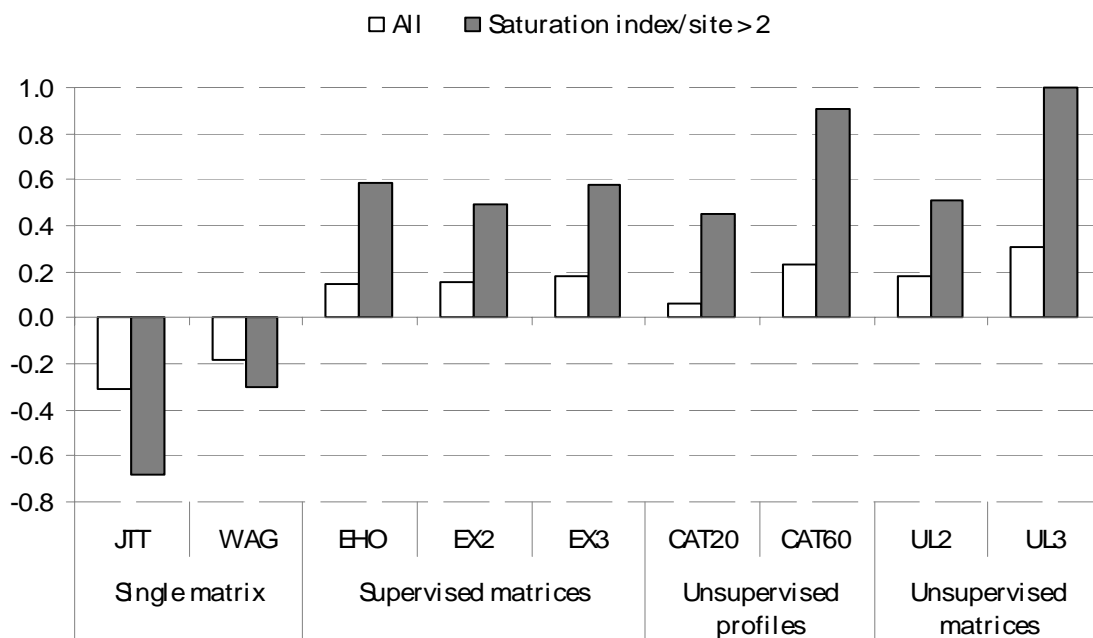
**Table 3: Unsupervised mixture model matrices**

Note:  $\mathbf{\Pi}_\theta/\text{SUP}$  : best correlation value of amino-acid frequencies between given matrix and supervised matrices from EX2 (first line, small letters) and EHO (second line, capital letters). For example, with M2, “0.969 b 0.842 E” means that amino-acid frequencies of M2 have correlations of 0.969 and 0.842 with those of buried sites and extended sites, respectively; moreover, correlations with other site categories are lower than these values.  $\mathbf{R}_\theta/\text{SUP}$  : same as  $\mathbf{\Pi}_\theta/\text{SUP}$  but using the log values of exchangeabilities. See note to Table 1 for other symbols.

	JTT	WAG	EHO	EX2	EX3	CAT20	CAT60	UL2	UL3
tree length	0.98	0.90	1.04	1.13	1.03	1.10	1.14	1.04	1.08
$\alpha$	0.99	1.17	1.07	1.19	1.17	0.94	0.97	1.12	1.21
#diff	35 (29)	40 (29)	34 (16)	36 (22)	35 (19)	49 (17)	48 (20)	37 (16)	38 (22)
R&F	0.17	0.20	0.13	0.13	0.14	0.26	0.26	0.14	0.15

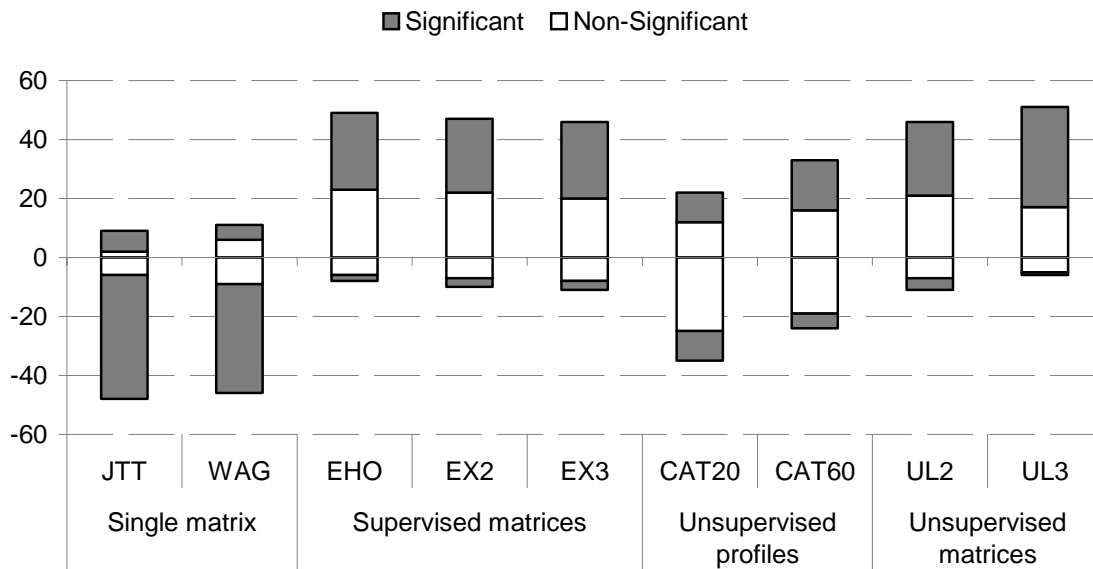
**Table 4: Model comparison, regarding tree length, gamma shape parameter and topology**

Note: These results are obtained with the 57 TreeBase test alignments. The tree length is the sum of branch lengths,  $\alpha$  denotes the gamma shape parameter, and the Robinson and Foulds (1981) topological distance corresponds to the number of clades that belongs to one tree but not the other. This distance is normalized and ranges between 0 (both trees are identical) and 1 (they do not share any clade in common). Symbols are as follows: tree length: average of the ratios between given model and LG tree lengths;  $\alpha$ : average of the ratios between given model and LG  $\alpha$  values; #diff: numbers of alignments where given model and LG topologies differ (numbers between parentheses count the significant differences using the KH test with  $p < 0.01$ ); R&F: average of the Robinson and Foulds distance between given model and LG topologies.



**Figure 1: AIC/site gain compared to LG**

Note: All models are compared to LG. Negative gains (JTT and WAG) mean that the models are worse than LG, while positive gains correspond to (mixture) models that improve LG. The gains are provided for all 57 TreeBase test alignments, and for the 8 alignments with saturation index per site larger than 2.



**Figure 2: number of alignments with better/worse likelihood values than LG**

Note: Number of alignments (among the 57 TreeBase test alignments) where each model provides a better (positive side) and a worse (negative side) likelihood value than LG. The grey bars correspond to the numbers of significant differences using the Kishino-Hasegawa test with  $p < 0.01$ .