An improved general amino acid replacement matrix

Quang S. Le, Olivier Gascuel

An improved general amino-acid replacement matrix

Si Quang LE & 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, gascuel@lirmm.fr

* Corresponding author

Abstract

Amino-acid replacement matrices are an essential basis of protein phylogenetics. They are used to compute substitution probabilities along phylogeny branches, and thus the likelihood of the data. They are also essential in protein alignment. A number of replacement matrices and methods to estimate these matrices from protein alignments have been proposed since the seminal work of Dayhoff et al. (1972). An important advance was achieved by Whelan and Goldman (2001), who designed an efficient maximum-likelihood estimation approach that accounts for the phylogenies of sequences within each training alignment. We further refine this method by incorporating the variability of evolutionary rates across sites in the matrix estimation, and using a much larger and diverse database than BRKALN, which was used to estimate the WAG matrix. To estimate our new matrix (called LG), we use an adaptation of the XRATE software and 3912 alignments from Pfam, comprising ~50,000 sequences and ~6.5 million residues overall. To evaluate the LG performance, we use an independent sample consisting of 59 alignments from TreeBase, and randomly divide Pfam alignments into 3,412 training and 500 test alignments. The comparison with WAG and JTT shows a clear likelihood improvement. With TreeBase, we find that: (1) the average AIC gain per site is 0.25 and 0.42, when compared to WAG and JTT, respectively; (2) LG is significantly better than WAG for 38 alignments (among 59), and significantly worse with 2 alignments only; (3) tree topologies inferred with LG, WAG and JTT frequently differ, indicating that using LG impacts the likelihood value but also the output tree. Results with the test alignments from Pfam are analogous. LG and a PHYML implementation can be downloaded from http://atgc.lirmm.fr/LG.

Keywords: amino-acid substitutions; replacement matrices; JTT ; WAG ; maximum-likelihood estimations; phylogenetic inference.

Introduction

Amino-acid replacement matrices are 20×20 matrices which contain estimates of the instantaneous substitution rates from any amino acid to another one. Let $\mathbf{Q} = \left(q_{xy}\right)$ be such a matrix and assume that **Q** accurately models the substitution process; the probability $p_{xy}\left(dt\right)$ of observing a change from amino-acid *x* to amino-acid *y* $\left(x \neq y\right)$ during a short period of time dt is equal to $q_{xy}dt$, while the probability $p_{xx}\left(dt\right)$ that amino-acid *x* is unchanged equals $1 + q_{xx}dt = 1 - \sum_{y \neq x} q_{xy}dt$. The rates in **Q** reflect the biological, chemical and physical properties of amino acids, e.g. replacements between arginine (positively charged) and aspartate (negatively charged) are under negative selection and have low rate, while replacements between isoleucyne and valine (both hydrophobic, aliphatic and very non-reactive) are frequent and have high rate (see textbooks, e.g. Betts and Russell 2003). Amino-acid replacement matrices are essential for inferring 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 change probabilities along the tree branches, and thus the likelihood of the data (see textbooks, e.g. Felsenstein 2003, Bryant et al. 2005, Yang 2006). Moreover, replacement matrices are closely related to score matrices, which are essential for aligning proteins and computing alignment scores (see textbooks, e.g. Setubal and Meidanis 1997). Applications of protein evolution models are reviewed in (Thorne 2000).

A number of replacement matrices and estimation methods have been proposed since the seminal work of Dayhoff et al. (1972). The first approaches exploited the linearity between p_{xy} probabilities and q_{xy} rates with low timescale. They considered closely related sequence pairs (typically with >85% identity), counted the number of amino-acid changes of each type per pair, rescaled these change numbers based on the sequence divergence for the analyzed pair, and averaged the results for all sequence pairs (for details see, e.g. Setubal and Meidanis 1997, Kosiol and Goldman 2004). The popular Dayhoff (1978) and JTT (Jones et al. 1992) matrices were estimated using this counting approach. A drawback of this type of method is that only closely related sequence pairs can be used. If the threshold is too low, a number of sequence pairs are discarded from the analysis. When the threshold is too high, linearity is no longer ensured due to the presence of hidden substitutions. This limitation was alleviated by Müller and Vingron's (2000) resolvent method, which can exploit more diverging sequence pairs than simple counting methods. Matrix logarithm (instead

of linear interpolation) is also used for the same purpose in numerous replacement matrix estimation methods (Benner et al. 1994, Arvestad and Bruno 1997, Devauchelle et al. 2001, Veerassamy et al. 2003, Arvestad 2006). However, all of these methods (simple counting, resolvent, or logarithmic) are only able to deal with independent sequence pairs, and cannot exploit multiple alignments and the corresponding phylogenies, so substantial evolutionary information is overlooked.

Adachi and Hasegawa (1996), Yang et al. (1998) and Adachi et al. (2000) first attempted to benefit from multiple alignments using maximum-likelihood (ML) approaches. Due to the computational burden, they used relatively restricted datasets (less than 100,000 residues) composed of concatenated protein alignments from a few species (20, 23 and 10, respectively). All proteins were assumed to share the same phylogeny, which was estimated by ML simultaneously with the replacement matrix. Whelan and Goldman (2001) showed that approximate phylogenies can be used to obtain accurate matrix estimates, thus considerably simplifying the computations by avoiding simultaneous optimization of the replacement matrix, phylogenies and branch lengths. They used a much larger database than in previous ML studies, BRKALN (D. Jones, unpublished data), containing 182 alignments and ~900,000 residues. They first inferred the phylogenies using NJ (Saitou and Nei 1987), re-estimated the branch-lengths by ML under the JTT model, and estimated the optimal replacement matrix by ML using an expectation-maximization (EM) algorithm. Their WAG matrix showed a clear improvement over JTT and Dayhoff matrices with respect to the likelihood values of inferred phylogenies (see also our results).

A second way to improve amino-acid replacement modeling is to use different matrices depending on the data or sequence sites. Replacement matrices have been estimated for various domains of life (e.g. Dimmic et al. 2002, Abascal et al. 2006), organelles (e.g. Adachi and Hasegawa 1996, Adachi et al. 2000), protein types (e.g. Jones et al. 1994) or protein families (e.g. Arvestad 2006). Replacement matrices have also been estimated for various site categories, mostly based on the solvent accessibility and secondary structure (e.g. Koshi and Goldstein 1995, Thorne et al. 1996, Goldman et al. 1998, Holmes and Rubin 2002, Lartillot and Philippe 2004).

Here, we present a new general amino-acid replacement matrix. 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), with a very large number of alignments from the three kingdoms of life. In fact, general matrices are still widely used, even though specific matrices should be preferred for certain analyses, e.g. with membrane or mitochondrial proteins. Moreover, the method we propose to

estimate our general matrix should be effective to obtain dedicated matrices for special protein groups or site classes. This method is a continuation of Whelan and Goldman's (2001) ML estimation procedure but, unlike this latter method, it incorporates the variability of rates across sites in likelihood calculations and replacement rate estimations. It is commonly acknowledged that sites of a given protein do not all evolve at the same rate, i.e. some sites are slow (or invariant) due to strong functional or structural constraints, while other sites with low evolutionary pressure (usually situated in turns) evolve rapidly. The standard approach (Yang 1993) to account for among-site rate variations in ML tree inference involves using a discrete gamma distribution of rates (+Γ option), which is often combined with a category of invariant sites (+I option; Gu et al. 1995). Using these options greatly increases the tree likelihood, and most of the phylogenies that are published today have been inferred under models that account for among-site rate variation. We show that the same holds for replacement matrix estimation, and that accounting for site rates enhances estimations of the rates of amino-acid changes. Moreover, to estimate our new matrix, we use a much larger and diverse dataset than BRKALN, which was used to estimate the WAG matrix.

In the following, we first describe our data, then our estimation method. We compare our results with JTT and WAG, and, finally, discuss directions for further research.

Datasets

To estimate our replacement matrix, we use Pfam (Bateman et al. 2002), which contains an extensive collection of protein families and domains. Overall, the current version (May 2007) of Pfam contains 6,885 Pfam families that match ~75% of protein sequences in Swiss-Prot and TrEMBL (Boeckmann et al. 2003). We use the seed alignments of Pfam, which are manually verified multiple alignments of representative sets of sequences corresponding to each Pfam family. To avoid learning from too restricted alignments, we first run GBLOCKS (Castresana 2000; default options) to eliminate sites containing many gaps, and then select all alignments with at least 5 taxa and 50 (remaining) sites. We thus obtain 3,912 alignments (49,637 sequences, 599,692 sites and 6,697,813 residues), with few gaps (~1% of the residues) and sufficient numbers of taxa (~13 per alignment, on average) and sites (~153 per alignment, on average). These alignments have several relevant properties to estimate a general replacement matrix: (1) they are highly diverse, as they represent (through manual selection) more than half of the Pfam families; (2) they are high quality alignments (thanks to manual curation); (3)

they contain a moderate number of taxa, which facilitates computations, notably the inference of phylogenetic trees prior to matrix estimation.

The BRKALN database (186 alignments, 3,905 sequences, 50,867 sites and 895,132 residues) which was used to estimate the WAG matrix is more restricted than ours and contains some very large alignments (up to 100 taxa), but also some very small ones (2 taxa). Moreover, protein families were included in BRKALN only if the 3-D structure of at least one member of the family was experimentally determined when the database was built (mid-1990s). This likely induces some bias toward specific globular proteins, which are typically easy to crystallize with well-defined 3-D structure.

To avoid estimating and testing our replacement matrix with the same data, which could overrate its real performance, we randomly select 500 alignments for testing from the 3,912 (cleaned for gaps and large enough) Pfam alignments, thus leaving 3,412 alignments for training.

Moreover, to check that our matrix is not biased in favor of Pfam alignments, we also use test alignments from TreeBase (Sanderson et al. 1994). This database contains alignments that have been used for phylogenetic purposes and deposited on the database by the author prior to publication. Thus, most of these 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% of gapped sites; some alignments are very large (up to ~12,500 sites), while some others are limited (minimum of 7 and 55 taxa and sites, respectively). All protein alignments from TreeBase (May 2007) are selected, except 3 of them because the set of taxa differs in the alignment and in the published tree, and 2 of them because the maximum pairwise divergence seems excessively large in a phylogenetic inference context (>2.0 substitutions per site, using a standard WAG distance). Moreover, 5 redundant alignments are removed. We thus obtain 59 test alignments, among which 2 correspond to genomic data with concatenated protein sequences. The average number of sequences per alignment is ~25, and the average number of sites is ~550 for non-genomic alignments, and is above 12,500 for genomic ones. These alignments are larger than Pfam alignments and should be representative of usual phylogenetic studies. All our test alignments are downloadable from http://atgc.lirmm.fr/LG.

Model and estimation method

We assume (as usual) a general time-reversible model of amino-acid substitutions. We first describe this model, its components and use to infer phylogenies, then our method to estimate this model from protein alignments.

The general time-reversible substitution model and its use in tree inference

This section provides notation and the main properties. More details and explanations can be found in textbooks (e.g. Felsenstein 2003, Bryant et al. 2005, Yang 2006).

We assume that sites evolve independently, and that the substitution process is time-continuous and -homogeneous. The evolution of any given site is characterized by a Markovian substitution matrix, which is denoted $\mathbf{Q} = \left(q_{xy}\right)$ and remains constant during evolution. The set of states corresponds to the 20 amino acids, q_{xy} $(x \neq y)$ is the substitution rate between amino-acids *x* and *y*, and q_{xx} diagonal terms are such that the row sums are all zero.

Moreover, the evolutionary process is assumed to be stationary. The stationary (or equilibrium) distribution is denoted $\mathbf{\Pi} = \left(\pi_x\right)$, where π_x is the probability of amino-acid *x*. $\mathbf{\Pi}$ and $\mathbf{Q}$ are dependent $\left(\mathbf{\Pi Q} = 0\right)$, and the empirical distribution of amino acids within the dataset being studied should be close to $\mathbf{\Pi}$.

Finally, the process is assumed to be time reversible. We use this property to rewrite $\mathbf{Q} = \left(q_{xy}\right)$ as

$$\begin{aligned} q_{xy} &= \pi_y r_{x \leftrightarrow y}, \; x \neq y, \\ q_{xx} &= -\sum_{y \neq x} q_{xy}, \end{aligned} \tag{1}$$

where $\mathbf{R} = \left(r_{x \leftrightarrow y}\right)$ is symmetric, independent of $\mathbf{\Pi}$, and is called the exchangeability matrix. Equation (1) is commonly used (F option, available in several programs) to adapt the $\mathbf{Q}$ matrix to proteins with an atypical amino-acid distribution; we simply multiply the exchangeability coefficients $\left(r_{x \leftrightarrow y}\right)$ by the amino-acid frequencies $\left(\pi_y\right)$ in the studied proteins.

In molecular phylogenetics, times and branch lengths are measured in number of substitutions per site rather than years. Thus we normalize $\mathbf{Q}$ so that a time unit ($t = 1.0$) corresponds to 1.0 expected substitution per site. The normalized form $\dot{\mathbf{Q}}$ of $\mathbf{Q}$ is defined by:

$$\dot{\mathbf{Q}} = \frac{1}{\mu}\mathbf{Q} = \left(\frac{q_{xy}}{\mu}\right), \text{ with normalization term } \mu = -\sum_x \pi_x q_{xx} . \quad (2)$$

In the following, we shall estimate non-normalized **Q** matrices, in order to have more flexibility in rate estimation; such a matrix can be written as $\mathbf{Q} = \rho\dot{\mathbf{Q}}$, where ρ is a global rate. However, normalized matrices will be used in tree inference, as usual, and **Q** will denote a normalized matrix unless explicitly stated.

Amino-acid changes over the course of time are represented by the matrix $\mathbf{P}(t) = \left(p_{xy}(t)\right)$, where $p_{xy}(t)$ is the probability of observing a change from *x* to *y* when the elapsed time is *t*. Note that hidden substitutions are possible, and that $p_{xy}(t)$ sums all possibilities (1, 2, 3 … *n* … substitutions, with an initial state *x* and final state *y*). As stated above, 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 (**Q**) and change probabilities (**P**):

$$\mathbf{P}(t) = e^{\mathbf{Q}t}, \quad (3)$$

where the right term denotes the matrix exponential.

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

$$L(T, \mathbf{Q}; D) = \prod_i L(T, \mathbf{Q}; D_i), \quad (4)$$

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* (D_i) given *T* and **Q**. $L(T, \mathbf{Q}; D_i)$ is computed by applying Equation (3) 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. The most common way to account for this fact is to assume that rates vary across sites and follow a gamma distribution (Yang 1993). Moreover, as in most datasets some sites are constant (i.e. contain a single amino acid), the gamma model is usually improved when assuming that some sites are invariant and do not undergo any substitution along the studied phylogenetic tree (Gu et al. 1995). Practical implementation of these assumptions relies on discrete categories of rates. Each site belongs to a category $c \in \{1, 2, ...C\}$, with probability π_c and rate ρ_c. Yang's approach involves categories with equal probabilities (i.e. $\pi_c = 1/C$) and ρ_c rates being defined by parameter α of the gamma distribution. When accounting for invariant sites, we have a special category (with

zero rate) and one more parameter, denoted π_{invar}, which corresponds to the proportion of invariant sites. Note that the proportion of invariant sites is always lower than that of constant sites in the dataset, as some constant sites may have undergone hidden substitutions. Altogether, the likelihood of the data for tree *T*, replacement matrix **Q**, and gamma distributed rate categories with invariant sites, is:

$$L(T,\mathbf{Q},\alpha,\pi_{invar};D)=\prod_i\left[\pi_{invar}L(Invariant;D_i)+(1-\pi_{invar})\sum_{1\le c\le C}\frac{1}{C}L(T,\rho_c\mathbf{Q};D_i)\right], \quad (5)$$

where: $L(D_i;Invariant)$ is the likelihood of site *i* assuming the invariant model, i.e. 0 if the site is non-constant, or else π_x when the site is constant and contains amino-acid *x*; $\rho_c\mathbf{Q}$ is simply the matrix of rates from **Q** multiplied by the category rate ρ_c, i.e. $\rho_c\mathbf{Q}=(\rho_c q_{xy})$. As the rate for invariant sites is null, we also have $L(invariant;D_i)=L(T,0\times\mathbf{Q};D_i)$. Moreover, it is easily seen from Equation (3) that Equation (5) can be rewritten as:

$$L(T,\mathbf{Q},\alpha,\pi_{invar};D)=\prod_i\left[\pi_{invar}L(Invariant;D_i)+(1-\pi_{invar})\sum_{1\le c\le C}\frac{1}{C}L(\rho_c T,\mathbf{Q};D_i)\right], \quad (6)$$

where $\rho_c T$ is the same as tree *T*, but with all branch lengths being multiplied by ρ_c.

The standard approach to infer trees from protein sequences is to search for the tree *T* (with branch lengths) which maximizes likelihood (4) or (6), assuming that amino-acid substitutions are modeled by a given replacement matrix **Q** (WAG, JTT, etc.). Model parameters, i.e. π_{invar} and α that defines ρ_c rates, are usually estimated along the way, as they vary from one dataset to another. When the F option is turned on, we estimate the equilibrium frequencies of amino acids using the empirical frequencies in the dataset, or by likelihood maximization; otherwise we use the default frequencies of the **Q** matrix at hand. Estimating amino-acid frequencies involves 19 additional free parameters to be accounted for in Akaike (1974), BIC (Schwartz 1978) and related criteria (Posada and Buckley 2004).

Estimating the replacement matrix from alignments

We now have a set of protein alignments, denoted $A=\{D^a\}$, where D^a is an alignment, and we aim to estimate the **Q** matrix using a maximum-likelihood approach. The likelihood of *A* is

$$L(A)=\prod_a L(T^a,\mathbf{Q};D^a), \quad (7)$$

where T^a is a phylogenetic tree relating the sequences in D^a, and the product runs over all alignments in *A*. However, maximizing likelihood (7) is a hard task since we should both maximize **Q** and the T^a trees, i.e. a huge number of numerical parameters (**Q** coefficients and branch lengths) plus the tree topologies. Whelan and Goldman (2001) greatly simplified the computations using a two-step approach. They first estimated an approximate tree T^a for every alignment D^a, and then used these trees in Equation (7). Based on this simplification, the likelihood of *A* becomes

$$L(A) = \prod_a L\left(\mathbf{Q}; D^a, T^a\right), \tag{8}$$

that is, the likelihood of the data and of the T^a trees, given replacement matrix **Q**. Using likelihood (8), only **Q** coefficients have to be estimated. The rationale of this simplification is based on the common observation that estimates of evolution model parameters (typically, the gamma shape parameter) remain relatively constant across near-optimal tree topologies. This is assumed to be the case for parameters used to describe amino-acid replacement. Notably, relative values of amino-acid exchangeability parameters $\left(r_{i\leftrightarrow j}\right)$ are assumed to stay approximately constant over near-optimal branch lengths and tree topologies, and the global rate ρ is used to fit the exchangeabilities to the branch lengths in the current trees. Thus, as soon as the T^a trees in Equation (8) are sufficiently close (within a scaling factor ρ) to the optimal trees, the estimations of **Q** from (7) and from (8) should be nearly identical. This was checked in (Whelan and Goldman 2001) by iterating the optimization process; almost no change was observed when T^a trees were refitted to a first estimation of **Q** and **Q** was reestimated using these improved trees. To obtain reasonably accurate T^a trees, Whelan and Goldman (2001) used the Neighbor Joining algorithm with Dayhoff distances, and reestimated the branch lengths by maximum likelihood with JTT and the F option. No gamma distribution of rates and no invariant sites were used in these tree estimations, nor were they used in **Q** estimation, where Equation (4) was used in Equation (8) to compute the likelihood of each alignment.

However, since rates do vary across sites, the **Q** matrix is not optimally estimated using this method. For example, the constant sites (~18% of the sites in our Pfam alignments) have a strong influence on **Q** estimation, whereas they are likely invariant and do not provide much information on amino-acid substitutions. In the same way, highly variable sites are not properly accounted for in **Q** estimation because their changes along the tree are mostly explained by their high substitution rate, rather than by the detailed features of the replacement process.

We use the following estimation procedure to account for rates across sites:

(a) Just as in (Whelan and Goldman 2001), likelihood (8) is used to estimate **Q**. However, the T^a trees are inferred by maximum-likelihood with rates across sites; for each alignment D^a, PHYML (Guindon and Gascuel 2003) is run with WAG, the invariant site model and 4 discrete categories of gamma distributed rates (i.e. WAG+Γ4+I). The gamma shape parameter (α) and the proportion of invariant sites (π_{invar}) are estimated from the data (D^a).

(b) For every site *i*, the posterior probability of each rate category is computed using

$$\begin{aligned} &\frac{(1-\pi_{invar})}{C} L\left(\rho_c T^a, \mathbf{Q}; D_i^a\right) \quad \text{for the discrete gamma categories, and} \\ &\pi_{invar} L\left(invariant; D_i^a\right) \qquad \text{for the invariant category.} \end{aligned} \tag{9}$$

Let $c(i)$ be the rate category with maximum posterior probability (9) for site *i*, and $\rho_{c(i)}$ the corresponding rate ($\rho_{c(i)} = 0$ when $c(i)$ refer to the invariant category). To compute the likelihood of any D^a alignment and T^a tree in Equation (8), a simplified version of Equation (6) is used, i.e.

$$L\left(\mathbf{Q}; D^a, T^a\right) = \prod_i L\left(\mathbf{Q}; D_i^a, \rho_{c(i)} T^a\right). \tag{10}$$

The difference with respect to Equation (6) is that the weighted sum over all rate categories is replaced with the most likely category, thus highlighting its most likely evolutionary rate for each site. In other words, Equation (10) does not integrate over site categories and represents the likelihood of the data, T^a trees and selected site categories, given replacement matrix **Q**. When using Equation (6) in likelihood (8), the results are not any better than those obtained with Equation (10) and we are faced with the presence of numerous local optima of the likelihood function. Moreover, likelihood (10) induces a slight bias when using a category of invariant sites. Indeed, a large proportion of constant sites (~40%, typically containing the most conserved residues, e.g. proline) are classified in the invariant category and are not accounted for in the matrix estimation, since the site likelihood assuming the invariant model does not depend on **Q**. Several approaches are possible to deal with this difficulty, as detailed in the Appendix. Based on computer simulations and the results obtained with our test sets, it appears that the best solution is not to use any invariant category. All constant sites are thus retained in rate estimations and are classified in the slowest gamma category, and their (tiny) influence is accounted for. This simple approach is equivalent to classify all sites in the gamma rate category with maximum posterior probability (9) and divide every alignment into 4 sub-alignments,

with each containing the sites that belong to a given category c with rate ρ_c ($\neq 0$). Note that the ρ_c rates are not the same from one alignment to another, as they depend on the gamma shape parameter separately estimated by PHYML for each alignment.

(c) The sub-alignments and their associated rescaled trees ($\rho_c T^a$) are independent. The estimation task is thus equivalent to that induced by Whelan and Goldman's (2001) approach and their use of Equation (4) to compute the likelihood of each alignment. We simply have 4 times more alignments and associated trees, but the same total number of sites, which is the key factor as computing likelihood (8) basically involves running over all the sites. Whelan and Goldman (2001) designed an expectation-maximization (EM) algorithm to solve **Q** estimation. We use XRATE (Holmes and Rubin 2002, Klosterman et al 2006), which is a powerful and flexible EM tool to estimate replacement matrices and other more complex probabilistic models. As all EM approaches, XRATE is faced with local optima and is sensitive to starting values. Thus, we initialize XRATE with WAG exchangeability matrix, the starting amino-acid equilibrium distribution is set at the empirical frequencies in our training set, and the starting global rate ρ equals 1.0. Moreover, we use the forgiven option (with 3 jumps) to escape from local optima. XRATE requires about 6 to 8 hours on our cluster (16 X 2.33GHz biprocessors with 8 Go RAM) to process our (huge) training set. Experiments with other starting points, including random matrices, indicate that XRATE performs remarkably well as all output matrices have nearly identical likelihood values (the difference is $\sim 10^{-3}$ log-likelihood points per site, while the average log-likelihood value per site is nearly -20.0), but also that the optimization task is a difficult one as all matrices are different and correspond to local optima or to a large, flat region of the likelihood surface.

(d) However, the best way to improve the replacement matrix is to iterate the learning process, rather than spending computing time on matrix optimization with fixed trees and site categories. Let LG1 be the matrix inferred using the above procedure, then we again run steps (a, tree inference), (b, site classification) and (c, matrix estimation using XRATE), but replace WAG by LG1 in all of these steps. The results shown in the Appendix indicate that the second iteration matrix, called LG2, is slightly but noticeably better than LG1. However, starting from LG2 and again running the learning procedure, XRATE is unable to improve LG2, which is a (local) optimum with respect to Equation (7). Moreover, LG2 remains nearly identical when all 3,912 (training plus testing) Pfam alignments are used to learn again starting from LG2**.** This indicates that sampling differences (3,912 alignments versus 3,412 alignments) do not markedly affect the rate estimates. Moreover, our results do not

contradict the assumption in Equation (8) that nearly optimal trees are sufficient to obtain accurate rate estimates; rates in LG2 and LG1 are highly correlated (0.997 when using the log-values) and LG2 is only a refinement of LG1. LG2 is thus our final LG matrix, which is available from http://atgc.lirmm.fr/LG.

The properties of this estimation procedure (e.g. presence of bias, benefit obtained by iterating the learning process) are further studied and discussed in the Appendix.

Results

We first describe the main features of the thus-estimated LG matrix, and then compare its performance in tree inference to several other replacement matrices with different options and datasets.

LG replacement matrix

As stated above, the LG matrix (as estimated using the above procedure) is defined by 3 components: the global rate (ρ), the amino-acid equilibrium distribution (Π), and the exchangeability matrix (**R**). We describe each of these components in turn.

The global rate (ρ) is equal to 1.11 and 1.07 for the first (LG1) and second (LG2) iterations, respectively. This indicates that LG is globally faster than WAG, but it is difficult to extrapolate the LG properties from these findings. To study the LG rate in tree inference, we thus measure the tree length obtained with the normalized version of LG and with WAG, both used with 4 gamma categories and invariant sites. The results are displayed in Table 1 for Pfam and TreeBase test alignments. This table also provides a comparison between LG and WAG regarding the estimate of the gamma shape parameter (α). These results highlight a clear difference between LG and WAG: LG trees are ~10-15% longer on average than WAG trees, and this finding is observed with almost all test alignments. We also observe that the variability of rates among sites is higher (α is lower) with LG than with WAG, and, again, this is observed with most alignments. Both findings are consistent as evolutionary distances and branch lengths are increased when the α value decreases. We shall see that LG trees also tend to be more likely than WAG trees. All of this means that LG better characterizes the evolutionary patterns than WAG, and thus captures more hidden substitutions, which results in longer trees (see Pagel and Meade, 2005, for a discussion on tree length and likelihood value).

Figure 1 displays the amino-acid frequencies (Π) of: (1) WAG (i.e. that observed in BRKALN), (2) our Pfam alignments, and (3) LG. All values are highly correlated, though we see some differences, e.g. with glycine (~8% and ~6% with WAG and LG, respectively). Moreover, Pfam and LG frequencies (obtained by ML estimation, starting from Pfam values) are very close but not identical, e.g. with glycine (~7% and ~6% with Pfam and LG, respectively). We shall see that these moderate differences in amino-acid equilibrium frequencies do not explain the differences between WAG and LG in tree inference.

In fact, most differences between WAG and LG are explained by their exchangeability matrices (**R**). Figure 2 provides a bubble plot representation of the exchangeability coefficients in WAG and LG. A quick visual inspection reveals that WAG and LG coefficients are highly correlated; coefficients that are high (low) in one matrix are also high (low) in the other one. Basically, the two matrices describe similar biological, chemical and physical properties of the amino acids. However, looking (Figure 3) at the relative differences between WAG and LG, we see that these two matrices are quite different. Some LG coefficients are up to ~6 times lower than the corresponding WAG coefficients. In fact, LG coefficients are much more contrasted than WAG coefficients. For example, the fastest and slowest coefficients correspond to the same amino-acid pairs in WAG and LG (isoleucine↔valine and cystein↔glutamic-acid, respectively), but the fastest/slowest ratio equals ~3000 for LG and ~350 for WAG. Looking at the sum of the 10 fastest and 10 slowest (again amino-acid pairs are pretty much the same for WAG and LG), the ratio becomes ~375 for LG and ~100 for WAG. The same holds when comparing LG with JTT and with WAG' (learned from our Pfam alignments when using the same estimation procedure as WAG, see below), so the high LG contrast is induced by our estimation procedure, rather than by our training alignments. Basically, this procedure is better able to distinguish among substitution events that are very rare (likely occurring in fast sites only) and those that are not so rare (possibly occurring in slow sites). Significantly, 10 among the 14 substitutions that require 3 changes at the codon level have lower coefficients in LG than in WAG, e.g. cystein↔lysine with 0.013 and 0.078 for LG and WAG, respectively. Within the 4 exceptions, 2 have nearly the same coefficients in LG and WAG, and the 2 others correspond to relatively large exchangeability values, e.g. cysteine↔methionine with 0.894 and 0.410 for LG and WAG, respectively (for recent results on codon substitutions, see Kosiol et al. 2007). Frequent events are also better characterized as the rate of the sites where they occur is accounted for. However, LG cannot simply be viewed as a contrasted version of WAG, e.g. arginine↔tryptophane and cysteine↔threonine both have relatively high coefficients, but the former has 0.593 and 1.221 in LG

and WAG, respectively (ratio ≈ 0.5), while the latter has 1.143 and 0.538 in LG and WAG, respectively (ratio ≈ 2.0). Careful inspection, analysis and interpretation of LG exchangeabilities would thus be deserved.

Performance comparisons

Our aim is to assess the performance of LG when used to estimate the likelihood of phylogenies and to infer them from data. LG is compared with 3 replacement matrices using various options. These are standard JTT and WAG matrices, and a new matrix, denoted WAG', which we learn from our Pfam training alignments using Whelan and Goldman (2001) procedure (i.e. using Equations (4) and (8), instead of Equations (8) and (10) for LG; see above for details and Appendix for further comparisons). Thus, WAG' measures the gain (relative to WAG) brought by using Pfam instead of BRKALN, and the difference between LG and WAG' represents the gain obtained when using our estimation method, in comparison with that of Whelan and Goldman. Note that all performance comparisons are based on the test alignments, which are independent of the training alignments used to learn LG (and WAG').

Unless explicitly stated (i.e., -Γ-I), all models are used with 4 discrete gamma rate categories and invariant sites (i.e. +Γ4+I, not written for conciseness). Most models are used both with and without the F option. When the F option is turned on (i.e. +F), the empirical amino-acid frequencies in the alignment are used in Equation (1) to adapt the replacement matrix to the specificities of the analyzed dataset. Otherwise, the F option is turned off (i.e. -F, implicit when no indication is provided), and we use the default amino-acid frequencies of the model. Better likelihood values are expected when the F option is turned on, but the +F option requires estimation of 19 additional free parameters (amino-acid frequencies) which may counterbalance in the AIC criterion the likelihood gain in comparison with -F. Moreover, with Pfam test alignments, we expect WAG and JTT to be closer to WAG' and LG when used with the +F option, than when used with -F, as with +F all models use the same (Pfam-like) amino-acid frequencies. Finally, we also test WAG but with LG amino-acid frequencies (denoted WAG+LGF) to measure the relative impact of the amino-acid frequencies and exchangeability matrices. Having WAG similar to WAG+LGF (in likelihood value) would mean that most of the difference between WAG and LG is induced by their exchangeability matrices.

All models and options are run on the 500 (Pfam) and 59 (TreeBase) test alignments using PHYML with standard options. The starting tree is built by BIONJ (default option of PHYML), and

we perform an SPR search of the tree space (Hordijk and Gascuel 2005). The gamma shape parameter (α) and the proportion of invariant sites (π_{invar}) are estimated from the data.

For all models and options, we measure the AIC criterion (Akaike 1974) on each of the test alignments, i.e.

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

where: $LL\left(M, T^a; D^a\right)$ is the log-likelihood of alignment D^a given model M and inferred tree T^a; $\#\,parameters\left(M\right)$ is the number of parameters of model M. All tested models induce one parameter (length) per tree branch, plus 2 parameters with the +Γ4+I option, plus 19 parameters with the +F option. We also define the average AIC per site of model M for the alignment dataset A, which is simply

$$AIC/site\left(M, A\right) = \sum_a AIC\left(M, D^a\right) \Big/ \sum_a s^a , \qquad (11)$$

where s^a is the number of sites in D^a. All models are compared to WAG and its variants, using criterion (11). To complete this global average result, we also count the number of alignments in A where $AIC\left(M_1, D^a\right) > AIC\left(M_2, D^a\right)$, where M_1, M_2 is any model pair. Moreover, to assess the statistical significance of the observed difference between models M_1 and M_2, we use the non-parametric paired sign test (MacStewart 1941). For every alignment D^a, we compare the number of *positive* sites $\left(LL\left(M_1, T_1^a; D_i^a\right) > LL\left(M_2, T_2^a; D_i^a\right)\right)$ and the number of *negative* sites $\left(LL\left(M_1, T_1^a; D_i^a\right) < LL\left(M_2, T_2^a; D_i^a\right)\right)$. When the number of positive sites is significantly larger than the number of negative sites (p-value < 0.01) and when $\left(LL\left(M_1, T_1^a; D^a\right) > LL\left(M_2, T_2^a; D^a\right)\right)$, we say that M_1 is significantly better than M_2 with alignment D^a. Inferred trees T_1^a and T_2^a can be identical or different. However, this test applies only to models having the same number of parameters since there is no penalty for the parameter number as in AIC. This (quite simple) test is a non-parametric version of the Kishino-Hasegawa (KH; 1989) test. It avoids any normality assumption (as in several KH-test versions), and is fully applicable here as there is no selection bias that would favor one model compared to the other (for more explanations see Felsenstein 2003; Goldman et al. 2000). We selected this version of KH because it emphasizes the number of sites that prefer model M1 over M2, which seems better suited for model comparison than relying on the high effect of few sites (typically highly unlikely, with strongly negative log-likelihood values).

Finally, we also compare the topology of inferred trees and count the number of alignments where the tree built using any model *M* is not the same as the tree inferred with WAG (or one of its variants). The true tree is not known with real data (as opposed to simulated data), and our aim is to measure the impact of the various models in terms of topology, i.e. Do we frequently infer a different tree topology when improving the replacement matrix? Indeed, it is commonly believed that tree topologies inferred with usual matrices (WAG, JTT, etc.) tend to be identical, which would mean that any efforts to refine these matrices are somewhat useless. When different topologies are found, we should prefer the one with best likelihood value, which is likely inferred using an accurate replacement matrix with relevant model options. However, the difference may be slight and non-significant, so we cannot reject the topology with the lower likelihood value. Thus we also count the number of cases where the *M* and WAG topologies differ, and where the difference in AIC value between the two models is statistically significant (using the sign test, see above).

All comparisons are displayed in Table 2 (59 TreeBase test alignments) and Table 3 (500 Pfam test alignments). Figure 4 also shows the progress of the various models compared to JTT. We first discuss differences among models in terms of AIC values, and then the topological impact.

AIC values

Tables 2 and 3, and Figure 4 are congruent. We see that:

- WAG-Γ-I is (as expected) a poor model. With TreeBase WAG-Γ-I is never better than WAG (+Γ4+I), and this occurs for only 26 datasets among the 500 Pfam alignments. These 26 datasets are limited (~7 taxa and ~100 sites on average), which penalizes the 2 additional parameters required by +Γ4+I option. Moreover, the AIC gain per site of WAG-Γ-I is quite low (~0.0006) when averaged within the 26 files. Similar results are found with JTT and LG (see additional results on the LG website). Clearly, among-site rate variation has to be accounted for in phylogenetic inference, as already discussed in a number of papers. However, adding invariant sites (+Γ4+I) improves only a little (see LG web site) compared to using gamma distributed rates (+Γ4-I).
- WAG+F slightly improves WAG with TreeBase, and the gain is a bit higher with Pfam. However, the number of datasets that are better with WAG+F than with WAG is pretty much the same as the number of datasets that are worse, and this holds both with TreeBase and Pfam. Moreover, when using the BIC criterion (Schwartz 1978), WAG+F is worse than WAG, both

with Treebase (BIC difference per site equals -0.04) and Pfam (BIC difference per site equals -0.28). The detailed analysis (not shown) indicates that, as expected, the datasets having a large number of sites tend to be improved using the +F option, while small datasets are penalized by the 19 additional parameters (amino-acid frequencies) to be estimated with +F.

- WAG+LGF slightly improves WAG. This is expected with Pfam as LG amino-acid frequencies are estimated from this database. However, the gain is low (0.01 with TreeBase and 0.04 with Pfam) and not significant with TreeBase. This shows that the difference between WAG and LG is not (or marginally) induced by their amino-acid frequencies, but rather by their exchangeabilities.
- JTT and JTT+F are clearly worse than WAG and WAG+F, respectively, both with TreeBase and Pfam, and the difference is significant for a number of alignments. Similar observations were provided by Whelan and Goldman (2001) with BRKALN. Their results are confirmed here with independent alignments.
- LG clearly improves WAG, with an average AIC gain per site of 0.25 and 0.21 for TreeBase and Pfam, respectively. With an alignment length of 300 (standard value for proteins), the expected gain of LG over WAG is about 70-75 AIC points, which is equivalent to 35-40 log-likelihood points as WAG and LG have the same number of parameters. For most alignments AIC value of LG is better than that of WAG, both with TreeBase (48 among 59) and Pfam (409 among 500). Moreover, the LG gain is often significant (38 and 161 times for TreeBase and Pfam, respectively), while WAG is rarely significantly better than LG (2 and 6 times, respectively). The 9 alignments that are better with WAG than with LG are of limited size and/or have a large number of gaps, and thus contain a low phylogenetic signal. The detailed analysis (see LG website) shows that the LG gain over WAG tends to increase when the variability of rates among sites increases, i.e. when the value of the gamma shape parameter (α) decreases. This is an expected result as LG is designed to cope with among site rate variation. But this is only a minor effect and for standard α values (say $\alpha < 4.0$, i.e. for >90% of datasets) LG is clearly better than WAG.
- LG+F also improves WAG+F in terms of average gain per site (0.20 and 0.13 for TreeBase and Pfam, respectively), but to a lesser extent than LG versus WAG. This is an expected result with Pfam, as WAG+F uses Pfam amino-acid frequencies instead of the BRKALN frequencies of standard WAG. LG+F is clearly better than WAG+F for a number of alignments, e.g. with TreeBase LG+F is 27 (among 59) times significantly better than WAG+F, while WAG+F is

better than LG+F with 7 datasets only. However, LG+F is not any better than LG. With AIC, the difference between these two models is nearly null, both with TreeBase and Pfam (as can be noted in Tables 2 and 3 when comparing with WAG and WAG+F). With BIC, LG+F is significantly worse than LG (not shown). However, just as with WAG, LG+F tends to be better than LG with large alignments.

- The comparison between LG and WAG' illustrates the gain provided by our estimation procedure, as WAG' is estimated from the same Pfam alignments as LG, but using Whelan and Goldman's (2001) approach. LG is clearly better than WAG', even if the average gain in AIC value is lower than with standard WAG (0.12 and 0.06 with TreeBase and Pfam, respectively). With TreeBase, about half of the gain between LG and WAG is explained by our estimation procedure, while with Pfam the proportion is about a third. The difference between TreeBase and Pfam (half versus a third) simply comes from the fact that WAG' closely fits (i.e. better than WAG) Pfam alignments since it is estimated from Pfam. Moreover, LG is better than WAG' for most alignments and the difference is very often significant, e.g. 50 times (among 59) with TreeBase. Finally, WAG' tends to be better than LG (see LG website) when variation in rates among sites is not used (option -Γ-I). Again, this is an expected result as LG accounts for variation in rates among sites, while WAG' does not. But both LG-Γ-I and WAG'-Γ-I are poor models (see LG website) that should be discarded in most analyses. Thus, in practice, LG outperforms WAG', and this result is obtained thanks to our estimation method.

All of these findings are summarized in Figure 4. WAG is a clear improvement over JTT, thanks to ML estimation of rates. WAG+LGF is slightly better than WAG, but the difference is mostly visible with Pfam as LG amino-acid frequencies are estimated from Pfam. WAG' is better than WAG+LGF since it is estimated (using the WAG procedure) from our large and diverse sample of Pfam alignments, which impacts both the exchangeabilities and amino-acid frequencies. LG is clearly best, thanks to its improved exchangeability matrix that is estimated by accounting for the site rates. It is worth noting that the difference between LG and WAG is even larger than that between WAG and JTT. Note, moreover, that the average results of Figure 4 are statistically significant for a number of alignments.

Tree topologies

Tables 2 and 3 show that changing the substitution model changes the inferred topology with ~1/2 (Pfam) to ~2/3 (TreeBase) of the alignments. As expected, the LG topology tends to have a higher

likelihood value than the WAG topology, while the JTT topology tends to be worse than that of WAG, e.g. with TreeBase, the LG topology is more likely than that of WAG with 31 alignments (among 40, where LG and WAG topologies differ), while the WAG topology is better than that of JTT with 26 alignments (among 36). A large number of these differences are statistically significant. Notably, with TreeBase, the LG topology is significantly better than that of WAG 25 times, and worse with 2 alignments only. Results obtained with Pfam alignments are less marked as the topological differences between LG, WAG and JTT are statistically significant with only 10% to 15% of the alignments. However, the main conclusions are unchanged, e.g. the LG topology is significantly better than that of WAG with 61 alignments and worse with 4. The difference between Pfam and TreeBase comes from the alignment size. Pfam alignments are somewhat limited, which is handy for rate matrix estimation, but induces low likelihood gains in a test context. In fact, when looking at the 100 largest alignments in our Pfam test set (~30 taxa and ~242 sites, on average), we find that the LG topology is significantly better than that of WAG with 37 alignments and is never significantly worse (among 67 alignments, where LG and WAG topologies differ). Thus, it appears that using LG (instead of WAG or JTT) impacts the tree topology with a large proportion (say half) of the alignments, and that these changes tend to be significant and in favor of LG when the alignments are sufficiently large.

Discussion

We propose in this paper an improved maximum-likelihood method to estimate amino-acid replacement matrices. This method accounts for among-site rate variation and provides accurate replacement rate estimates, as amino-acid changes observed in the data are rescaled depending on whether they occur in slow or fast sites. This method is used to estimate a general replacement matrix using a large, diverse and high-quality set of alignments extracted from Pfam. Our LG matrix shows marked differences with WAG. Most notably, the lowest exchangeabilities (typically corresponding to three substitutions at the codon level) are much lower in LG than in WAG. We tested the performance of LG, WAG and JTT using 59 alignments from TreeBase and 500 independent Pfam alignments. Our results show that LG outperforms WAG (itself outperforming JTT) with respect to the likelihood value of inferred trees. Most notably, LG is often significantly better than WAG, but very rarely worse. LG also tends to produce topologies that differ with respect to those of WAG, so LG should be preferred to WAG and JTT in a number of phylogenetic analyses.

There are several directions for further works. It would be relevant to study the properties and performance of LG in aligning proteins, as it was developed and studied in a pure phylogenetic context. Our estimation method could possibly be improved by using a standard expression of tree likelihood, i.e. by replacing our Equation (10) by the usual Equation (6). However, one is then faced with multiple local optima of the likelihood function and a hard optimization problem. In fact, even if XRATE and its EM-based approach perform remarkably well, we believe that an interesting direction for further works would be to search for faster and/or more robust estimation algorithms, e.g. based on other optimization principles. Finally, and most importantly, our estimation method could be applied to a number of datasets to obtain amino-acid replacement matrices specific to certain protein groups (e.g. intracellular, extracellular and membrane proteins), life domains (e.g. viruses, bacteria or apicomplexa), or structural configurations of sites (e.g. buried versus exposed to solvent).

Acknowledgements

Sincere thanks to Nicolas Galtier, Nick Goldman, Stéphane Guindon, Ian Holmes, Nicolas Lartillot and Simon Whelan for their help, suggestions and comments. This work was supported by ACI IMPBIO (ModelPhylo project) and ANR BIOSYS (MitoSys project).

References

Abascal F, Posada D, Zardoya R. 2007. MtArt: a new model of amino acid replacement for Arthropoda. Mol. Biol. Evol. 24(1):1-5.

Adachi J, Hasegawa M. 1996. Model of amino acid substitution in proteins encoded by mitochondrial DNA. J. Mol. Evol. 42:459–468.

Adachi J, Waddell PJ, Martin W, Hasegawa M. 2000. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. J. Mol. Evol. 50:348–358.

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

Arvestad L. 2006. Efficient Methods for Estimating Amino Acid Replacement Rates. J. Mol. Evol. 62:663–673.

Arvestad L, Bruno WJ. 1997. Estimation of reversible substitution matrices from multiple pairs of sequences. J. Mol. Evol. 45:696–703

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/

Benner S, Cohen M, Gonnet G. 1994. Amino acid substitution during functionally constrained divergent evolution of protein sequences. Protein Eng. 7:1323–1332.

Betts MJ, Russell RB. 2003. Amino acid properties and consequences of subsitutions. In: Barnes MR, Gray IC, editors. Bioinformatics for Geneticists. Wiley. Ville pages

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I et al. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 31:365-370.

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

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

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

Dayhoff M, Schwartz R, Orcutt B. 1978. A model of evolutionary change in proteins. In: Dayhoff MO, editor. Atlas of protein sequence and structure. Volume 5, Suppl. 3. National Biomedical Research Foundation, Washington, DC. p 345-352.

Devauchelle C, Grossmann A, Hénaut A, Holschneider M, Monnerot M, Risler J, Torrésani B. 2001. Rate matrices for analyzing large families of protein sequences. J. Comput. Biol. 8:381–399

Dimmic MW, Rest JS, Mindell DP, Goldstein RA. 2002. rtREV: a substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. J. Mol. Evol. 55:65-73.

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.

Goldman N, Thorne JL, Jones DT. 1998. Assessing the Impact of Secondary Structure and Solvent Accessibility on Protein Evolution. Genetics 149: 445–458.

Goldman N, Anderson JP, Rodrigo AG. 2000. Likelihood-Based Tests of Topologies in Phylogenetics. Syst. Biol. 49(4):652–670.

Gu X, Fu YX, Li WH. 1995. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. Mol. Biol. Evol. 12:546–557.

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.

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.

Kosiol C, Goldman N. 2004. Different versions of the Dayhoff rate matrix. Mol. Biol. Evol. 22 :193–199.

Kosiol C, Holmes I, Goldman N. 2007. An Empirical Codon Model for Protein Sequence Evolution. Mol. Biol. Evol. 24(7):1464–1479.

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.

MacStewart W. 1941. A note on the power of the sign test. Ann. Math. Statist. 12: 236-239.

Müller T, Vingron M. 2000. Modeling amino acid replacement. J. Comput. Biol. 7(6):761–776.

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.

Posada D, Buckley TR. 2004. Model selection and model averaging in phylogenetics: Advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst. Biol. 53:793-808.

Rambaut A and Grassly N. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:235-238.

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/

Schwarz G. 1978. Estimating the dimension of a model. Annals of Statistics 6: 461-464.

Setubal J, Meidanis J. 1997. Introduction to computational molecular biology. PWS Publishing Company, Boston.

Thorne J. 2000. Models of protein sequence evolution and their applications. Current Opinion in Genetics & Development 10:602–605.

Thorne JL, Goldman N, Jones DT. 1996. Combining Protein Evolution and Secondary Structure. Mol. Biol. Evol. 13:666-673.

Veerassamy S, Smith A, Tillier ER. 2003. A transition probability model for amino acid substitutions from blocks. J. Comput. Biol. 10:997–1010

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.

Appendix

In this appendix we discuss several variants of our estimation procedure and show, using simulations, that this procedure is nearly unbiased when a standard Markovian model of amino-acid replacement is assumed. We also provide comparisons to Whelan and Goldman's (2001) estimation method. Our aim is to study the properties of the approach (e.g. presence of bias, benefit obtained by iterating the learning process) and to check the influence of constant sites on rate estimations. Four variants of our estimation procedure are tested:

- MAP-Γ4: This is the method we describe in the core of the paper. All sites are classified in one of the 4 gamma categories, based on the maximum posterior probability (MAP, Equation (9)). As expected from simple considerations, all constant sites are classified in the slowest category.
- MAP-Γ4+I: All sites are classified into one of the 5 categories (4 gamma + 1 invariant) using MAP. As (constant) sites classified in the invariant category do not play any role in matrix estimation, they are removed from the training set. About 40% of the constant sites are discarded.
- NoConst: No constant site is incorporated in the training set, and the remaining (variable) sites are classified in one of the 4 gamma categories using MAP. This method exacerbates the (possible) bias induced by the invariant category. Constant sites represent about 18% of the sites in the original training set.
- RAND: for every site, we perform a random drawing of the category (among 5) based on the posterior probability (Equation (9)). This variant should reduce the bias (if any) as some constant sites with highly conserved residues (e.g. proline) will not be classified in the invariant category and will be incorporated in the training set. About 35% of the constant sites are discarded using this approach.

These 4 variants use the same basic parameters (T^a trees, α and π_{invar}). They are run with XRATE and require similar computing times (NoConst is a bit faster as it uses fewer sites than the other variants). Moreover, we also test the following method:

- NoRAS: this method is similar to Whelan and Goldman's (2001), with the only difference being that T^a trees are inferred using PhyML instead of NJ. This method does not use rates across sites (RAS) to infer T^a trees or in matrix estimation. It is run with XRATE and requires similar computing time as the others.

For each of these five estimation methods, we iterate the learning procedure, thus producing two matrices corresponding to the first and second learning steps. These matrices and estimation methods are compared using our test sets and simulated data. Simulated alignments are generated using SEQGEN (Rambault and Grassly 1997) and WAG+Γ4+I. The number (3,412) and sizes of these alignments are the same as in our training set. Moreover, the input trees and model parameters (gamma shape and invariant proportion) are those inferred by PhyML from the training set. This dataset thus mimics our training set, assuming that WAG+Γ4+I is the true evolutionary model, and the aim is to recover WAG from these data. We therefore run the 5 above methods using JTT to infer the T^a trees, compute the posterior probabilities of site rates and initialize the starting matrix in XRATE. We thus mimic (again) the estimation task with LG, which involves estimating a replacement matrix (i.e. LG with real data, and WAG with simulated data), knowing a reasonable approximation of this matrix (i.e. WAG with real data, and JTT with simulated data). As data are generated under a Γ4+I model, some bias is expected from the NoRAS procedure.

For each of the 5 estimation methods, each producing two matrices, we measure:

- The proportion of unexplained variance (PUV) in WAG, when learning from simulated data (*PUV in WAG* in Table 4). This criterion (standard in regression analysis) is computed here using the log-values of the matrix entries. Indeed, replacement rates are highly contrasted (see Results section) and using the rough values would only focus on the larger entries. Let $\mathbf{Q}_{ij}$ be an entry of the estimated **Q** matrix and $\mathbf{W}_{ij}$ the corresponding entry in WAG. The proportion of variance in WAG that is unexplained by **Q** is measured by

$$\mathrm{PUV}(\mathbf{Q}) = \frac{\sum_{i<j}\left(\log(\mathbf{W}_{ij}) - \log(\mathbf{Q}_{ij})\right)^2}{\sum_{i<j}\left(\log(\mathbf{W}_{ij}) - \overline{\log \mathbf{W}}\right)^2},$$

  where $\overline{\log \mathbf{W}}$ denotes the average log-values of the non-diagonal WAG entries. When $\mathrm{PUV}(\mathbf{Q})$ is null, then **Q** is identical to WAG. The larger $\mathrm{PUV}(\mathbf{Q})$, the worse is **Q** in estimating WAG. Due to the very large size of our simulated dataset, we have almost no sampling variance and PUV measures the bias of the estimation method.

- The correlation between the log-values of the non-diagonal entries of **Q** and of our final LG matrix (*LG correlation* in Table 4). **Q** and LG are learned from our Pfam training alignments (LG

is obtained using MAP-Γ4 and two learning steps). This criterion is used to measure the closeness of the different matrices estimated with non-simulated Pfam data.

- The difference in average AIC value per site between **Q**+Γ4+I and WAG+Γ4+I with TreeBase test alignments (*AIC per site TreeBase* in Table 4). **Q** is estimated using our Pfam training alignments. This criterion is the same as that used in Table 2.
- The difference in average AIC value per site between **Q**+Γ4+I and WAG+Γ4+I with Pfam test alignments (*AIC per site Pfam* in Table 4). Q is estimated using our Pfam training alignments. This criterion is the same as that used in Table 3.

All results are displayed in Table 4. The main conclusions are as follows:

- NoRAS is the worse method, both in terms of AIC per site and PUV (~8% of the variance in WAG is unexplained). This confirms that incorporating rates across sites in replacement matrix estimation is desirable. However, the AIC results are slightly better than those of WAG' (see Table 2 and 3) thanks to the use of PhyML (instead of NJ) T^a trees. In fact, the accuracy of the T^a trees used in the estimation procedure seems to be a critical parameter for this estimation method and the others. This explains why no method is able to fully recover WAG, despite the very large size of our training set; since some trees are erroneous due to small alignments, the task is simply impossible.
- The second iteration is a clear improvement over the first one, both in terms of AIC per site and PUV. The gain is not high but noticeable (except with NoRAS, for unclear reasons).
- The performance of the four variants of our estimation procedure is quite similar and the output matrices are all highly correlated (LG correlation > 0.995). As expected, we observe a slight bias with MAP-Γ4+I and NoConst when estimating WAG from simulated data (~1.5% of the variance in WAG is unexplained). But the AIC values of the four methods with real data are nearly the same, except NoConst which is slightly worse than the others with Pfam test alignments (but is best with TreeBase). This indicates that constant sites have a low impact on replacement matrix estimation, a finding which is easily understandable as they likely did not undergo any substitution within analyzed phylogenies.

Based on these observations, it is difficult to decide which variant is best. Both MAP-Γ4 and RAND seem to be nearly unbiased (~0.5% of the variance in WAG is unexplained) and obtain good AIC values. We decided to choose MAP-Γ4 in the core of the paper because the output matrix is

globally the best on our test sets. However, all four variants perform well and output similar matrices. Other experiments (not shown) indicate that our approach is robust, e.g. the results remain nearly identical when using 6 gamma categories (instead of 4), when the gamma shape parameter is the same for all training alignments, or when trees are inferred without invariant sites. Moreover, when using Equation (6) in place of Equation (10) to compute the tree likelihood (thanks to an appropriate phylo-grammar) the results remain similar, but the computing time is much increased and the EM-based estimation procedure of XRATE is faced with numerous local optima of the likelihood function.

		LG/WAG	#LG>WAG
Tree length	TreeBase	1.10	54
	Pfam	1.17	497
α	TreeBase	0.74	2
	Pfam	0.66	4

Table 1: Comparison of WAG and LG regarding the tree length and gamma shape parameter

Note: LG and WAG are run with PHYML using the Γ4+I option on TreeBase and Pfam test alignments. The tree length is the sum of all branch lengths; α denotes the gamma shape parameter; LG/WAG is the average of the ratios between LG and WAG values, over all alignments. #LG>WAG counts the number of alignments where the LG value is larger than the WAG value, among 59 and 500 alignments for TreeBase and Pfam, respectively. The sign test indicates that all these counts reveal highly significant differences between LG and WAG (p-value $\approx$ 0.0).

M1	M2	AIC per site	#M1>M2	#M1>M2 (p<0.01)	#M2>M1 (p<0.01)	#T1>T2	#T1>T2 (p<0.01)	#T2>T1 (p<0.01)
WAG-Γ-I	WAG	-1.39	0	-	-	0/41	-	-
WAG+F	WAG	0.04	27	-	-	16/35	-	-
WAG+LGF	WAG	0.01	26	6	5	14/36	3	5
JTT	WAG	-0.17	14	6	21	10/36	4	13
JTT+F	WAG+F	-0.22	12	5	41	8/32	3	18
LG	WAG	0.25	48	38	2	31/40	25	2
LG+F	WAG+F	0.20	46	27	7	31/39	16	6
LG	WAG'	0.12	51	50	1	32/37	31	1

Table 2: Model comparison with 59 test alignments from TreeBase

Note: All models (unless explicitly stated, i.e. -Γ-I) are run with PhyML using 4 categories of gamma distributed rates and invariant sites (i.e. Γ4+I option). +F involves using the empirical amino-acid distribution in the analyzed alignment, instead of the model default distribution. WAG+LGF has the same exchangeabilities as WAG, but uses the amino-acid frequencies of LG. WAG' is obtained using the WAG estimation procedure from our Pfam training alignments. Model M1 is compared to model M2 using 59 protein alignments from TreeBase. AIC per site: average per site difference in AIC value between M1 and M2; a positive (negative) value means that M1 is better (worse) than M2, on average. #M1>M2: number of alignments (among 59) where M1 has a better AIC value than M2. #M1>M2 (p<.01): number of alignments where the AIC of M1 is significantly better than that of M2; the paired sign test among sites is used to assess statistical significance, with p-value < 0.01 (this test does not apply when M1 and M2 do not have the same number of parameters). #M2>M1 (p<.01): same as #M1>M2 (p<.01), but now M2 is significantly better than M1. #T1>T2: number of alignments where the tree T1 inferred with M1 has a better AIC value than T2 inferred using M2, and where T1 and T2 have different topologies; this number is related to the total number of times where T1 and T2 have different topologies. #T1>T2 (p<.01): same as #T1>T2, but now T1 is significantly better than T2. #T2>T1 (p<.01): T2 is significantly better than T1.

M1	M2	AIC per site	#M1>M2	#M1>M2 (p<0.01)	#M2>M1 (p<0.01)	#T1>T2	#T1>T2 (p<0.01)	#T2>T1 (p<0.01)
WAG-Γ-I	WAG	-0.75	26	-	-	5/245	-	-
WAG+F	WAG	0.08	248	-	-	87/195	-	-
WAG+LGF	WAG	0.04	317	88	16	94/148	11	5
JTT	WAG	-0.19	97	15	127	37/198	3	42
JTT+F	WAG+F	-0.21	97	12	176	33/206	2	69
LG	WAG	0.21	409	161	6	180/214	73	2
LG+F	WAG+F	0.13	387	125	13	168/202	61	4
LG	WAG'	0.06	327	136	4	136/186	44	1

Table 3: Model comparison with 500 test alignments from Pfam

Note: See note to Table 2; all counts have to be referred to 500 alignments (instead of 59 with TreeBase in Table 2).

	Learning step	PUV in WAG	LG correlation	AIC per site TreeBase	AIC per site Pfam
MAP-Γ4	1st	0.017	0.9973	0.229	0.204
	2nd	0.007	1.0	0.246	0.208
MAP-Γ4+I	1st	0.023	0.9976	0.232	0.203
	2nd	0.012	0.9997	0.248	0.205
NoConst	1st	0.025	0.9977	0.235	0.200
	2nd	0.015	0.9985	0.248	0.195
RAND	1st	0.013	0.9952	0.224	0.203
	2nd	0.004	0.9979	0.243	0.209
NoRAS	1st	0.068	0.9868	0.157	0.167
	2nd	0.077	0.9875	0.158	0.169

Table 4: Comparison of 5 estimation methods with simulated (PUV) and Pfam (other columns) data

Note: PUV in WAG: proportion of unexplained variance in WAG; LG correlation: correlation with the final LG matrix that is selected in the paper; AIC per site TreeBase: difference in average AIC value per site with WAG using TreeBase test alignments; AIC per site Pfam: difference in average AIC value per site with WAG using Pfam test alignments. See Appendix for further explanations.

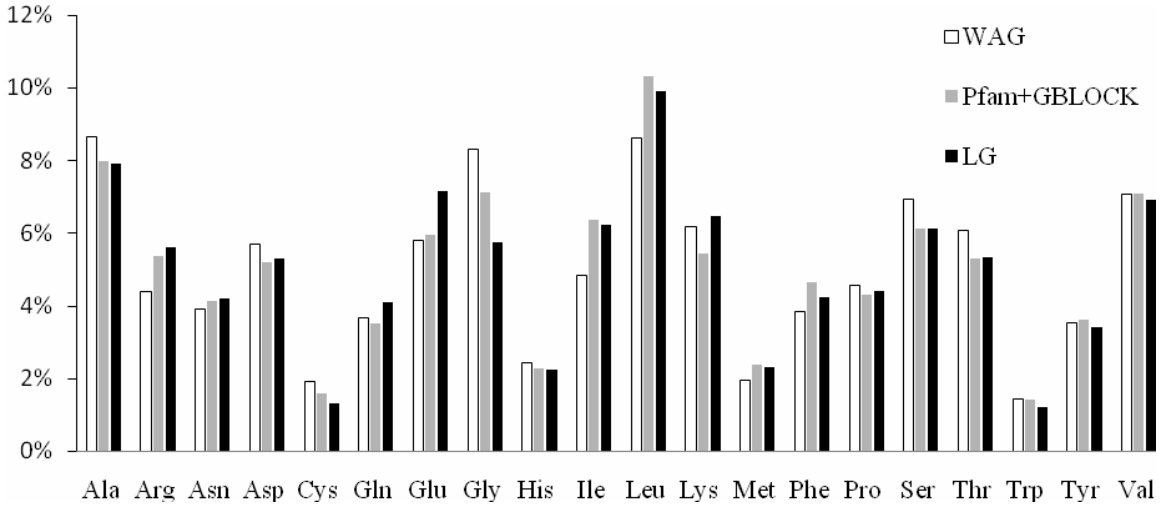

Figure 1: Amino-acid frequencies of WAG, Pfam alignments and LG

Note: The amino-acid frequencies of WAG correspond to those observed in the BRKALN database; Pfam+GBLOCK denotes our 3,412 training alignments cleaned for gaps with GBLOCK (see Datasets section); LG frequencies are obtained by ML optimization with XRATE (see Model and Estimation Method section).

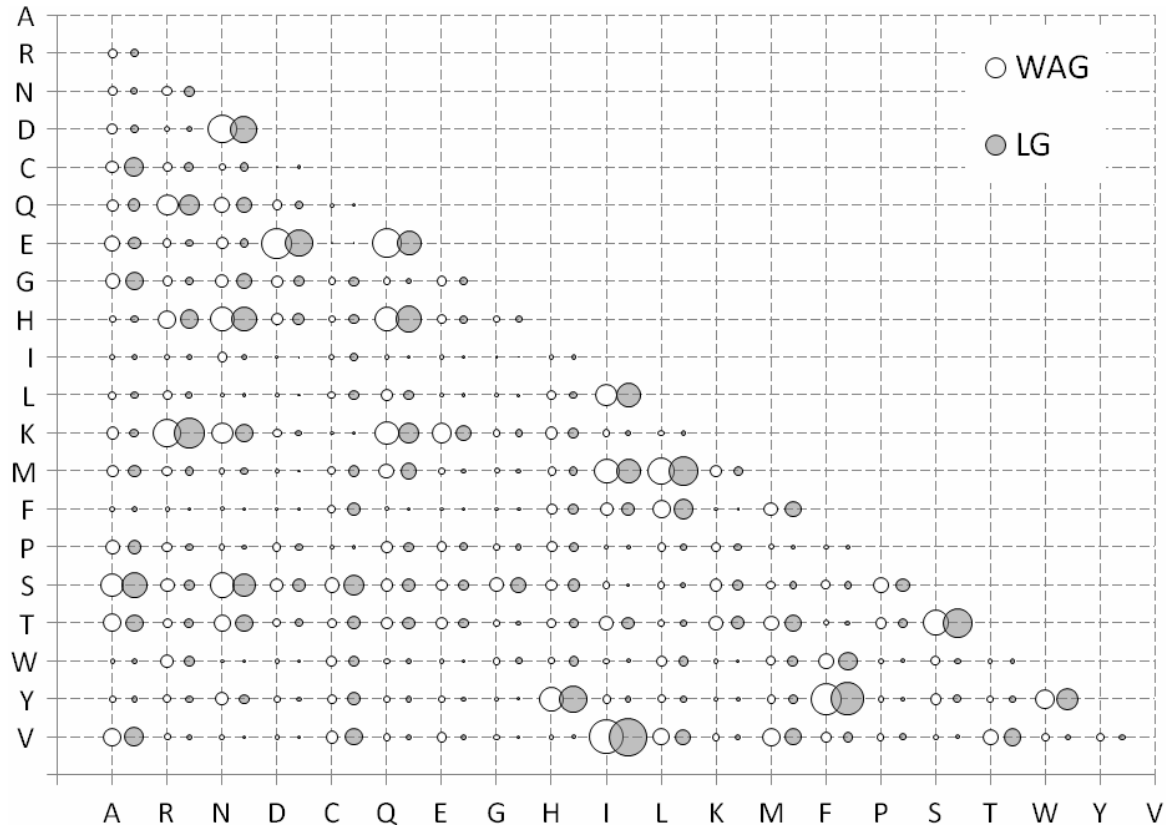

Figure 2: WAG and LG exchangeability coefficients.

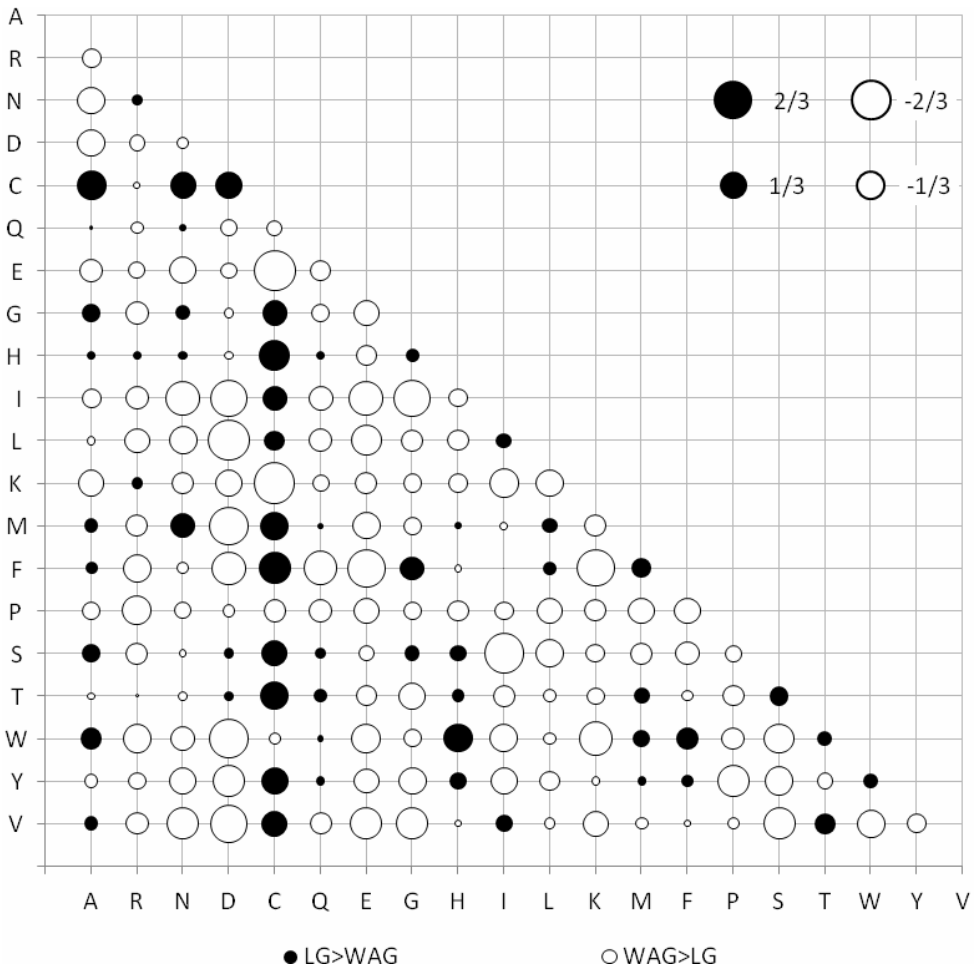

Figure 3: Relative differences between WAG and LG exchangeability coefficients.

Note: Each bubble represents the value of $\left(\mathrm{LG}_{ij} - \mathrm{WAG}_{ij}\right) \Big/ \left(\mathrm{LG}_{ij} + \mathrm{WAG}_{ij}\right)$, where M_{ij} denotes the exchangeability coefficient of matrix M between amino-acids *i* and *j*. Values of 1/3 and 2/3 mean that the LG coefficient is 2 and 5 times larger than that of WAG, respectively. -1/3 and -2/3 mean that WAG is 2 and 5 times larger than LG, respectively. The larger ratio corresponds to C↔F (cystein↔phenylalanine), where LG is ~2.6 times faster than WAG; the lower ratio corresponds to C↔E (cystein↔glutamic-acid), where LG is ~6.4 times slower than WAG.

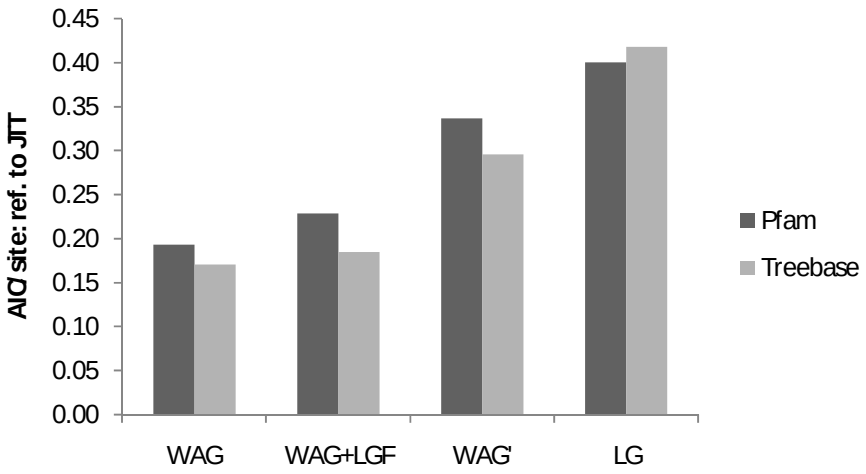

Figure 4: Progress in replacement matrix estimation compared to JTT.

Note: All models are run with PhyML using 4 categories of gamma distributed rates and invariant sites (i.e. Γ4+I option). Performance is measured by the average AIC per site and compared to the JTT value, e.g. the WAG gain over JTT with Pfam alignments is about 0.2 AIC point per site, meaning that with an alignment of 300 sites, the expected gain with WAG is about 60 AIC points, while the LG gain should be around 120 AIC points. As all these models have the same number of parameters, the difference in AIC value between two models is twice the difference in log-likelihood value. WAG+LGF has the same exchangeabilities as WAG, but uses the amino-frequencies of LG; WAG' is obtained using the WAG estimation procedure with our Pfam training alignments.