Application Note

ReplacementMatrix: a Web Server for Maximum Likelihood Estimation of Amino Acid Replacement Rate Matrices

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

[1]College of Technology, Vietnam National University Hanoi, 144 Xuan Thuy, Cau Giay, Hanoi, VIETNAM

[2]Méthodes et Algorithmes pour la Bioinformatique, LIRMM, CNRS – Université Montpellier 2, FRANCE

[3]Wellcome Trust Sanger Institute, Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Summary: Amino acid replacement rate matrices are an essential basis of protein studies (*e.g.* in phylogenetics and alignment). A number of general-purpose matrices have been proposed (*e.g.* JTT, WAG, LG) since the seminal work of Margaret Dayhoff and co-workers. However, it has been shown that matrices specific to certain protein groups (*e.g.* mitochondrial) or life domains (*e.g.* viruses) differ significantly from general average matrices, and thus perform better when applied to the data to which they are dedicated. This Web server implements the maximum-likelihood estimation procedure of Le and Gascuel (2008) and provides a number of tools and facilities. Users upload a set of multiple protein alignments from their domain of interest, and receive the resulting matrix by email, along with statistics and comparisons with other matrices. A non-parametric bootstrap is performed as an option, to assess the variability of replacement rate estimates. Maximum-likelihood trees inferred using the estimated rate matrix are also computed as an option, for each input alignment. Finely-tuned procedures and up-to-date ML software (PhyML 3.0, XRATE) are combined to perform all these heavy computations on our clusters.

Availability: http://www.atgc-montpellier.fr/ReplacementMatrix/

Contact: olivier.gascuel@lirmm.fr

1 INTRODUCTION

Amino-acid replacement matrices contain estimates of the instantaneous substitution rates from any amino acid to another. These rates reflect the biological, chemical and physical properties of amino acids. For example, we usually observe a high substitution rate between lysine (positively charged) and arginine (also positively charged) and a low substitution rate between lysine and aspartate (negatively charged). 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 closely related to score matrices, which are essential for aligning proteins and computing alignment scores.

Several general replacement matrices have been proposed such as PAM (Dayhoff *et al.* 1978), JTT (Jones *et al.* 1992), WAG (Whelan and Goldman 2001) and LG (Le and Gascuel 2008). These matrices were estimated from large and diverse sets of protein alignments. They tend to be robust and perform well in many cases. However, the performance of replacement matrices depends on life domains and protein groups (Keane *et al.* 2006). Replacement matrices have thus been estimated for specific domains (*e.g.* HIVw and HIVb matrices for HIV, Nickle *et al.* 2007; FLU for influenza, Dang *et al.* 2010) and proteins (*e.g.* mtREV for mitochondrial proteins, Adachi and Hasegawa 1996). It has been shown that often specific replacement matrices differ significantly from general matrices, and thus perform better when applied to the data to which they are dedicated (*e.g.* Adachi and Hasegawa 1996, Dang *et al.* 2010).

Since the seminal work of Dayhoff *et al.* (1978), a number of methods have been designed to estimate amino-acid replacement matrices from protein alignments. These methods belong to either counting (*e.g.* Jones *et al.* 1992) or maximum likelihood (ML) approaches (*e.g.* Adachi and Hasegawa 1996, Yang *et al.* 1998, Whelan and Goldman 2001). The former are limited to pairwise protein alignments, while the latter fully benefit from the information contained in multiple alignments and the corresponding phylogenies. Recently, we improved the ML method proposed by Whelan and Goldman (2001) by incorporating the variability of evolutionary rates across sites into the matrix estimation process (Le and Gascuel 2008). This procedure was successfully applied to estimate the LG matrix from 3,912 alignments of the Pfam database, the FLU matrix from 992 influenza protein alignments, and a number of matrices corresponding to different structural configurations of the residues (Le and Gascuel 2010).

The demand to estimate amino-acid replacement matrices for particular data is rising quickly because of the rapidly growing volume of sequence data and the ambition to better understand the evolution and relationships of specific protein groups and species. However, up-to-date replacement matrix estimation procedures are complex and highly demanding in computational terms. Our method (Le and Gascuel 2008) alternates tree building using PhyML (Guindon *et al.* 2010) and matrix estimation using XRATE (Klosterman *et al.* 2006), and involves complex data processing. It thus

*To whom correspondence should be addressed.

requires a huge amount of work to estimate a matrix from raw data sets. We describe here an implementation of this method in a web server. Users simply upload their alignments and receive by email the output matrix and a number of additional statistics and comparisons. In option, the server performs a non-parametric bootstrap to assess the variability of rate estimations, and infers the phylogeny of every input alignment using the estimated replacement matrix. In the following we first describe the estimation method and the bootstrap procedure, and then the web server, the input, the options and the various output files.

2 MODEL AND METHODS

The amino-acid substitution process is assumed to be independent among sites and lineages, and homogeneous during the course of evolution. The standard model is Markovian, time-continuous, time-reversible and represented by a 20 × 20 rate matrix $Q=\left(q_{ij}\right)$, where q_{ij} $(i \neq j)$ is the number of substitutions from amino acid *i* to amino acid *j* per time unit. The diagonal elements q_{ii} are such that the row sums are all zero. Any time-reversible matrix Q can be decomposed into a symmetric exchangeability matrix $R=\left(r_{ij}\right)$ and an amino-acid equilibrium frequency vector $\Pi=\left(\pi_i\right)$, using equality $q_{ij}=r_{ij}\pi_j$ $(i \neq j)$. Moreover, Q is normalized, that is $-\sum \pi_i q_{ii}=1$. Here we consider (as usual) the most general time-reversible model (GTR), which involves 189 (R) and 19 (Π) free parameters to be estimated from the data (see textbooks for additional explanation, *e.g.* Felsenstein 2003).

Given a set of protein alignments $D=\left\{D_a\right\}$, Q is estimated by maximizing the likelihood $L(D)=\prod L\left(T_a,\rho_a,Q;D_a\right)$, where the product runs over all alignments D_a and the inner term is the likelihood of D_a given the phylogenetic tree T_a, the rate across sites model ρ_a, and the replacement matrix Q. The rate across sites model used here is the standard discrete gamma distribution with 4 rate categories, and ρ_a represents the gamma distribution parameter associated with D_a.

Simultaneously optimizing T, Q and ρ parameters is computationally difficult. However, several authors showed that substitution model parameters (Q and ρ) can be accurately estimated using nearly optimal trees T. Whelan and Goldman (2001) estimated their WAG matrix by: (1) inferring the tree topologies using NJ, (2) estimating the tree branch lengths by ML assuming a JTT replacement process, and (3) estimating Q from the data and thereby inferred trees using a standard optimization procedure.

We refined this approach by incorporating an across-site rate model in the matrix estimation, namely 4 gamma categories plus invariant sites (Γ4+I). Our method (Le and Gascuel 2008) involves: (1) estimating the tree topologies and branch-lengths using PhyML (Guindon *et al.* 2010), (2) processing the alignment and trees to account for the rate model, (3) estimating Q from these processed data and trees using the expectation-maximization software XRATE (Klosterman *et al.* 2006), (4) iterating this procedure until $L(D)$ reaches a plateau. This estimation procedure is started using an approximate matrix. WAG was used to learn LG, and a nearly identical matrix was obtained when starting from JTT. We observed that 3 iterations are enough in practice, and that the invariant site category has little impact on Q estimation.

Above procedure is very heavy in computational terms. It is simplified here. The most time-consuming aspect is the ML estimation of trees, which is performed only once here (instead of ~3 times in the original procedure). Moreover, the rate model is simplified by using 4 gamma rate categories but no invariant sites (Γ4). The resulting matrix is nearly the same as that obtained using the full procedure (see results below) but the run time is 2-3 times faster. The simplified procedure is as follows:

1. Input a set of multiple alignments and a starting replacement matrix S; only exchangeabilities in S are used, frequencies are estimated from the data.
2. For each alignment, build a BioNJ tree and optimize the branch lengths and gamma rate parameter using PhyML with S and Γ4.
3. Process the alignments and trees to account for the Γ4 model (every alignment is divided into 4 sub-alignments using the posterior probability of site rate categories, and the 4 corresponding trees are rescaled using the rates estimated for each category under the gamma model).
4. Run XRATE with default options and S starting matrix to estimate a first matrix Q_1 from the processed alignments and trees (End of Step 1).
5. For each alignment, infer an ML tree using PhyML 3.0 with Q_1, Γ4 and the SPR tree search option.
6. Same as 3.
7. Same as 4, but replace S by Q_1 and output Q_2 (End of Step 2).
8. For each alignment, re-optimize the branch-lengths of the previously inferred ML tree and gamma rate parameter using PhyML with Q_2 and Γ4.
9. Same as 3 and 6.
10. Same as 4, but replace S by Q_2; output final Q matrix (End of Step 3).
11. For each alignment, re-optimize the branch-lengths of the previously inferred ML tree and the gamma rate parameter using PhyML with Q, with S, and with LG when $S \neq \mathrm{LG}$; output the corresponding log-likelihood and AIC values of every alignment and site for comparison purposes.

Only the second step in this procedure fully constructs an ML tree; the first step uses a distance-based tree topology (as with WAG estimation), while the third step reuses the ML topology inferred during the second step with a fairly accurate Q_1 matrix. Other parts are the same as in the original LG estimation procedure (except for the invariant site category, removed here).

When the final matrix has been estimated, it is sent to the user by email, along with a number of results, statistics and comparisons. Two additional options are available: (1) performing a bootstrap study to assess the variability of rate estimates; (2) running PhyML 3.0 with Q and standard options to infer the phylogenies estimated with the new matrix for all input alignments. These are expected to be significantly different from the phylogenies inferred with starting matrix S or LG. This option thus simultaneously estimates the replacement matrix and the trees, a task that cannot be achieved by any existing program, except some (*e.g.* Bayesian) when the input comprises a unique alignment. To save computing time, the starting trees and initial parameter values are taken from the above procedure.

The aim of the bootstrap procedure is to measure the variability of rate estimations depending on alignment selection. A large number of homogeneous alignments should provide reliable rate estimates, while a small data set and/or the use of heterogeneous alignments should result in poor estimations. Knowing the variability of rate estimates should be useful, for example, when studying and comparing the properties of amino acids in specific contexts (Kosiol *et al.* 2004), or when using replacement rate matrices in the search for non-standard genetic codes (Abascal *et al.* 2007).

The bootstrap procedure involves drawing with replacement $|D|$ alignments from D and running for each pseudo-sample the same estimation procedure as that used to estimate Q from D. This procedure is repeated a number of times to obtain an estimation of the distribution of the Q estimate. Because we draw alignments rather than individual sites, we obtain a measure of the sensibility of the estimated matrix to the choice of the alignments in D. We did not implement the alternative bootstrap procedure drawing sites rather than alignments, because most studies involve a very large number of sites (*e.g.* ~600,000 with LG estimation), which would result in over-confidence in estimated rate values.

The bootstrap procedure is highly time-consuming. We therefore perform only 10 replicates, thus obtaining 10 pseudo rate matrices from which we compute several statistics for each of the exchangeability $\left(r_{ij}\right)$ and frequency $\left(\pi_i\right)$ parameters. However, the procedure described above is still too heavy to be repeated 10 times. Thus, we re-use the same processed data and trees as in step 3, only running XRATE with the S starting matrix and the resampled set of alignments. Experimental studies show that these simplifications do not significantly affect the assessment of rate variability.

3 RESULTS

To illustrate the properties of the Web server, we re-estimated the LG matrix from the data set used in original publication (3,912 alignments, ~6 millions residues). We performed two runs of the bootstrap procedure, thus obtaining two estimates of the standard deviation for each of the π_i and r_{ij} parameters. The new matrix is nearly identical to the previous one (Pearson correlation ~0.9999). Regarding bootstrap results, we observed that: (1) the standard deviations associated to frequencies are quite small (deviation number $\pi_i/\sigma_i \approx 150$ on average); (2) the standard deviations associated with exchangeabilities are clearly larger ($r_{ij}/\sigma_{ij} \approx 40$ on average); (3) the relative difference between both standard deviations corresponding to the two bootstrap runs are moderate, despite the low number of replicates ($\left|\sigma_x^1 - \sigma_x^2\right| \Big/ \left|\sigma_x^1 + \sigma_x^2\right| \approx 0.15$ on average, for both π_i and r_{ij} parameters).

We performed the same experiment with 250 randomly selected alignments from the data set used to estimate the FLU matrix (~1.8 million residues). The new matrix is very close to the original one (Pearson correlation ~0.990), and we found the following average values for the previous measures: $\pi_i/\sigma_i \approx 48$; $r_{ij}/\sigma_{ij} \approx 14$; $\left|\sigma_x^1 - \sigma_x^2\right| \Big/ \left|\sigma_x^1 + \sigma_x^2\right| \approx 0.15$.

These results show that 10 bootstrap replicates are enough to obtain relevant measurements of the variability of estimations. We also note that the frequencies are accurately estimated with both data sets. In contrast, exchangeabilities are harder to estimate and show relatively high standard deviations, notably with FLU. Indeed, substitutions are partly hidden and substitution rates are not directly measurable on the sequences (as opposed to frequencies), especially for the amino-acid pairs that are rarely aligned together, as is common with highly conserved FLU alignments. Moreover, the bootstrap procedure resamples the alignments, rather than sites, and thus induces a large variability of estimations.

4 WEB SERVER, INPUT AND OUTPUT FILES

The main input is a set of multiple alignments in PHYLIP format. This typically contains hundreds or even thousands of alignments. However, each alignment must contain less than 100 sequences to reduce the computational burden. Larger alignments must be divided in several sub-alignments and given separately. A starting replacement matrix may also be provided, otherwise LG is the default. Two options allow for bootstrapping and running PhyML with the estimated matrix. The user receives an email with the estimated matrix along with a number of files and statistics. These include (see user guide for details):

- The new rate matrix in PAML triangular format, where exchangeability (r_{ij}) and frequency (π_i) parameters are given separately. These parameter values are compared to those of the starting matrix *S* and of LG (when $S \neq$ LG), using Pearson correlation, histograms and bubble graphs.
- A series of score matrices for various evolutionary distances (δ), derived from the rate matrix using standard log-odds: $log\left(\pi_i Pr\left(i \to j|\delta\right)\right)/\pi_i\pi_j\Big)$, where the probability of change from *i* to *j* given δ is computed by exponentiation of the rate matrix (Felsenstein 2003). As with PAM matrices, the δ distance ranges from 0.10 (corresponding to PAM10) to 2.5 (PAM250). These matrices can be used, for example, with MAFFT, CLUSTALW or BLAST to search for homologs or compute multiple alignments of specific protein groups.
- The fit of the new rate matrix to the input data is compared to that of *S* and of LG (when $S \neq$ LG), using the log-likelihood difference on the whole data set, divided by the total number of sites. To account for the fact that the new matrix is estimated from these data and thus has to be penalized for its (189+19) additional parameters, we use the AIC difference divided by the number of sites. The AIC and log-likelihood differences are also provided for every alignment and every site, for example to detect atypical alignments or site classes.

When the bootstrap and/or PhyML options have been checked (and confirmed after email reception of the first batch of results), the user receives separate emails containing the following:

- The standard deviation, the deviation number, the minimum and the maximum values (among 10 bootstrap estimates) for each of the frequency and exchangeability parameters.
- All trees inferred by PhyML 3.0 using the new matrix with SPR and standard options for each of the input alignments.

The current waiting time when all options are checked is ~12 days for very large Pfam data set, and ~4 days with FLU data set.

ACKNOWLEDGMENTS

Sincere thanks to Ian Holmes for his help with XRATE, and to the Vietnam National Foundation for Science and Technology Development and to ANR SYSBIO-MYTOSIS for financial support.

REFERENCES

Abascal, F. *et al.* (2007) MtArt: a new model of amino acid replacement for Arthropoda. *Mol. Biol. Evol.* **24**, 1–5.

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

Dang, C. *et al.* (2010) FLU, an amino acid substitution model for influenza proteins. *BMC Evol. Biol.* **10**, 99.

Dayhoff, M.O. *et al.* (1978) A model of evolutionary change in proteins. In: Dayhoff M.O. (ed) *Atlas of protein sequence and structure*. Vol. 5, Suppl. 3. *National Biomedical Research Foundation*, 345–352.

Felsenstein, J. (2003) *Inferring phylogenies*. Sinauer, Sunderland, MA.

Guindon, S. *et al.* (2010) New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321.

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

Keane, T.M. *et al.* (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.

Klosterman, P.S. *et al.* (2006) XRate: a fast prototyping, training and annotation tool for phylo-grammars. *BMC Bioinformatics* **7**, 428.

Kosiol, C. *et al.* (2004) A new criterion and method for amino-acid classification. *J Theor. Biol.* **7**, 97–106.

Le, S.Q. and Gascuel, O. (2008) An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307–1320.

Le, S.Q. and Gascuel, O. (2010) Accounting for Solvent Accessibility and Secondary Structure in Protein Phylogenetics is Clearly Beneficial. *Syst. Biol.* **59**, 277–287.

Nickle, D.C. *et al.* (2007) HIV-specific probabilistic models of protein evolution. *PLoS ONE* **2**, e503.

Whelan, S. and 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. *et al.* (1998). Models of amino acid substitution and applications to Mitochondrial protein evolution. *Mol. Biol. Evol.* **15**, 1600-1611.