BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 19 2011, pages 2758–2760
doi:10.1093/bioinformatics/btr435

Phylogenetics

Advance Access publication July 26, 2011

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 and Information Technology Institute, Vietnam National University, Hanoi, Vietnam, [2]Méthodes et algorithmes pour la Bioinformatique, LIRMM, CNRS – Université Montpellier 2, Montpellier, France and [3]Wellcome Trust Sanger Institute, Genome Campus, Hinxton, Cambridge CB10 1SA, UK

Associate Editor: David Posada

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 that was used to estimate LG, 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 optionally to assess the variability of replacement rate estimates. Maximum-likelihood trees, inferred using the estimated rate matrix, are also computed optionally for each input alignment. Finely tuned procedures and up-to-date ML software (PhyML 3.0, XRATE) are combined to perform all these heavy calculations on our clusters.

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

Contact: olivier.gascuel@lirmm.fr

Supplementary information: Supplementary data are available at http://www.atgc-montpellier.fr/ReplacementMatrix/

Received on March 1, 2011; revised on June 29, 2011; accepted on July 19, 2011

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. for HIV, Nickle *et al.*, (2007), and influenza, Dang *et al.* (2010)] and protein groups [e.g. mitochondrial proteins, Adachi and Hasegawa (1996)]. It has been shown that specific replacement matrices often 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 3912 alignments of the Pfam database, the FLU matrix from 992 influenza protein alignments and a number of matrices corresponding to various 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 a desire 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) involves complex data processing and alternates tree building using PhyML (Guindon *et al.*, 2010) and matrix estimation using XRATE (Klosterman *et al.*, 2006). It thus requires a huge amount of work to estimate a matrix from raw datasets. Here, we describe an implementation of this method in a Web server. Users upload their alignments and receive the output matrix by email along with a number of additional statistics and comparisons. Optionally, the server performs a non-parametric

*To whom correspondence should be addressed.

bootstrap to assess the variability of rate estimations, and infers the phylogeny of every input alignment using the estimated replacement matrix.

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 $Q = (q_{ij})$, 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 $R = (r_{ij})$ and an amino acid equilibrium frequency vector $\Pi = (\pi_i)$, 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 (GTR) model, 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 $D = \{D_a\}$, Q is estimated by maximizing the likelihood $L(D) = \prod L(T_a, \rho_a, Q; D_a)$, 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 site model ρ_a and the replacement matrix Q. Here we use the standard discrete gamma distribution with four rate categories, and ρ_a is the gamma parameter associated with D_a.

Simultaneously optimizing T, Q and ρ parameters is computationally difficult. However, several authors have 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: (i) inferring tree topologies using NJ; (ii) estimating tree branch lengths by ML assuming a JTT replacement process; and (iii) 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 four gamma categories plus invariant sites ($\Gamma 4+\mathrm{I}$). Our method (Le and Gascuel, 2008) involves: (i) estimating tree topologies and branch lengths using PhyML (Guindon *et al.*, 2010); (ii) processing alignment and trees to account for the rate model; (iii) estimating Q from these processed data and trees using the expectation–maximization software XRATE (Klosterman *et al.*, 2006); and (iv) 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 three iterations are enough in practice and that the invariant site category has little impact on Q estimation.

The above procedure is very heavy in computational terms. It is simplified here. The most time-consuming aspect is the ML estimation of tree topologies, which is performed only once here (instead of ~3 times in the original procedure). Moreover, the rate model is simplified by using four gamma rate categories, but no invariant sites ($\Gamma 4$). 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 has three main estimation steps (1, 2 and 3) and is as follows:

Step 0: input a set of multiple alignments and a starting replacement matrix S; only exchangeabilities in S are used, frequencies are estimated from the data.

Step 1: (a) For each alignment, build a BioNJ tree and optimize the branch lengths and gamma rate parameter using PhyML with S and $\Gamma 4$.

(b) Process the alignments and trees to account for the $\Gamma 4$ model: every alignment is divided into four subalignments using the posterior probability of site rate categories, and the four corresponding trees are rescaled using the rates estimated for each category under the gamma model.

(c) Run XRATE with default options and S starting matrix to estimate a first matrix Q_1 from the processed alignments and trees.

Step 2: (a) For each alignment, infer an ML tree using PhyML 3.0 with Q_1, $\Gamma 4$ and the SPR tree search option.

(b) Same as (1b).

(c) Same as (1c), but replace S by Q_1 and output Q_2.

Step 3: (a) For each alignment, re-optimize the branch lengths of the previously inferred ML tree and gamma rate parameter using PhyML with Q_2 and $\Gamma 4$.

(b) Same as (1b).

(c) Same as (1c), but replace S by Q_2 and output final Q matrix.

Step 4: 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 Step (2) in this procedure fully constructs an ML tree; Step (1) uses a distance-based tree topology (as with WAG estimation), while Step (3) reuses the ML topology inferred during Step (2) 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 returned along with a number of results, statistics and comparisons. Two additional options are available: (i) performing a bootstrap study to assess the variability of rate estimates; and (ii) running PhyML 3.0 with Q and standard options to infer the phylogenies estimated with the new matrix for all input alignments. When the latter option is used, the pipeline simultaneously estimates the replacement matrix and the trees from the input alignments. These are expected to be significantly different from the phylogenies inferred with starting matrix S or LG. To save computing time, the starting trees and initial parameter values are taken from Step (4) in the above procedure.

The aim of the bootstrap procedure is to measure the variability of rate estimations. This should be useful, for example, when 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 is performed in a standard manner: for every alignment D_a in D, we draw with replacement $|D_a|$ sites and then run the estimation procedure to obtain a pseudo rate matrix; this is repeated several times and the pseudo matrices are used to compute several statistics (e.g. the standard deviation) for each of the frequency (π_i) and exchangeability (r_{ij}) parameters. This procedure is highly time consuming, and we thus only perform 10 replicates. Moreover, the estimation scheme described above is still too heavy to be repeated 10 times. We therefore use the trees and site rate categories computed by PhyML with Q in Step (4), and run XRATE only once for each replicate, starting from the S matrix. Experimental studies show that these simplifications do not significantly affect the variability measures.

3 RESULTS WITH TWO SAMPLE DATASETS

To illustrate the properties of the Web server, we re-estimated the LG matrix from the data used in original publication (3912 alignments, ~6.5 millions residues) and the FLU matrix using 100 randomly selected alignments from the original dataset (~1.8 million residues). We performed a bootstrap with 500 (LG) and 1000 (FLU100) replicates to obtain accurate measures of the variability of parameter estimates, and 20 standard pipeline bootstrap runs with 10 replicates each. Detailed results are available as Supplementary Material from the Web site and summarized in Table 1.

We see that the new LG matrix estimated by the Web server is nearly identical to the published matrix, despite the simplifications in the estimation procedure. The new FLU100 matrix (estimated from 100 alignments) is also very close to

Table 1. Results with LG and FLU100 datasets

	R^2	σ_i/π_i	σ_{ij}/r_{ij}	$R^2(\sigma_i/\pi_i)$	$R^2(\sigma_{ij}/r_{ij})$
LG	0.996	0.004	0.044	0.81 ± 0.07	0.94 ± 0.01
FLU100	0.987	0.029	0.185	0.89 ± 0.03	0.88 ± 0.04

R^2: Pearson's correlation of the matrix estimated by the Web server and the published matrix. σ_i/π_i (σ_{ij}/r_{ij}): average relative deviation of the frequencies (exchangeabilities) obtained with 500 (LG) and 1000 (Flu100) bootstrap replicates. $R^2(\sigma_i/\pi_i)$ ($R^2(\sigma_{ij}/r_{ij})$): average and SD (among 20 trials) of the Pearson's correlations of the relative deviations of frequencies (exchangeabilities) computed with 10 replicates, and those computed with 500 (LG) and 1000 (FLU100) replicates.

the original one (estimated from 992 alignments). The relative deviations of equilibrium frequencies (σ_i/π_i) are quite low, while those of exchangeabilities (σ_{ij}/r_{ij}) are higher, especially with FLU100 where their average is nearly equal to 20%. This finding shows that exchangeabilities are much more difficult to estimate than frequencies. Exchangeabilities measure instantaneous rates of change, which are not directly observable from the data (as frequencies are) and may correspond to hidden changes between ancestral states. With highly conserved alignments as FLU100's, some amino acid pairs are rarely seen together in the same alignment site (e.g. Cys-Lys is present four times only among ~37 000 sites, see Supplementary Material), and thus estimating their exchangeabilities is inevitably a difficult task. Finally, we see that the bootstrap variability measures with 10 replicates are clearly correlated with those obtained using a large number of replicates (500 and 1000), and should thus be useful for analyzing the differences between rates or between matrices.

4 WEB SERVER, INPUT AND OUTPUT FILES

The main input is a set of multiple alignments in PHYLIP or Fasta format. This typically contains hundreds or even thousands of alignments. However, each alignment must contain <100 sequences to reduce the computational burden. Larger alignments must be divided into several subalignments that are 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 a job status URL and the estimated matrix by email 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 with those of the starting matrix S and of LG (when $S \neq$ LG), using Pearson's correlation, histograms and bubble graphs.
- A series of score matrices for various evolutionary distances (δ), derived from the rate matrix using standard log odds: $\log(\pi_i \Pr(i \to j|\delta)/\pi_i\pi_j)$, where the probability of change from i to j given δ is calculated 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 with that of S and of LG (when $S \neq$ LG), using the log-likelihood difference on the whole dataset, 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. Z-tests are used to assess the significance of AIC differences.

When the bootstrap and/or PhyML options are selected, the user receives separate emails providing:

- The SD, relative deviation, minimum and 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 selected is ~10 days for the very large LG dataset, and ~2 days with FLU100.

ACKNOWLEDGEMENTS

We thank Ian Holmes and Christophe Dessimoz for their help

Funding: Vietnam National Foundation for Science and Technology Development; French ANR,MITO-SYS project (BIOSYS06_136906).

Conflict of Interest: none declared.

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*. National Biomedical Research Foundation, Washington, DC, pp. 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. *Bioinformatics*, **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.