Accounting for Solvent Accessibility and Secondary Structure in Protein Phylogenetics is Clearly Beneficial

Si Quang LE\textsuperscript{1,2} and Olivier GASCUEL\textsuperscript{1*}\\
lsquang@gmail.com, gascuel@lirmm.fr

1: Méthodes et Algorithmes pour la Bioinformatique\\
LIRMM, CNRS - Université Montpellier II,\\
161 rue Ada, 34392 – Montpellier Cedex 5 – France

2: Wellcome Trust Sanger Institute,\\
Hinxton, Cambridge, CB10 1SA, United Kingdom

* Corresponding author

Online Appendix

1. GTR rate matrices

2. Experiments with PASSML

3. Comparison of CONF/LG, PART and MIX

See our web site (http://atgc.lirmm.fr/phyml-structure/) for:

Replacement rate matrices

PhyML implementation

Detailed results

Data
1. GTR rate matrices

All of our matrices comply with the general time-reversible (GTR) model. Such a matrix contains estimates of the instantaneous substitution rates from any amino acid to another one, and is denoted as $Q = (q_{xy})$, where $q_{xy}$ is the rate of replacement from $x$ to $y$ ($x \neq y$). $Q$ can be decomposed into three independent components using

$$q_{xy} = \rho \pi_y r_{x+y}, \quad x \neq y,$$

$$q_{xx} = -\sum_{y \neq x} q_{xy},$$

where: $\rho$ is the global rate of $Q$, equal to the expected number of substitutions per time unit; $\Pi = (\pi_x)$ is the vector of amino-acid equilibrium frequencies; $R = (r_{x+y})$ is the (symmetric) exchangeability matrix, which represents the general propensity of exchanges between amino-acids, independently of the amino-acid frequencies within the studied sequences (represented by $\Pi$). $R$ is normalized (i.e. $\rho = -\sum \pi_x q_{xx}$), and thus $Q$ contains $(\rho)+19(\Pi)+189(R)=209$ free parameters to be estimated from the data.

When a single replacement matrix is used (e.g. WAG), it is normalized (i.e. $\rho = 1$) to obtain a simple branch-length interpretation in terms of number of expected substitutions per site. When several site categories are used, matrices are no longer normalized (e.g. exposed sites evolve about thrice as fast as buried sites) and the model requires a specific global normalization to ensure branch-length interpretation. Let $C$ be any site category, $P_C$ the proportion of $C$, and $\rho_C$ the global rate of the replacement matrix associated to $C$. Since the $P_C$ proportions vary from one data set to another one, we do not normalize the model but rescale the inferred tree. This post-processing involves multiplying every branch length by the expected global rate (i.e. $\sum P_C \rho_C$), after all parameters (topology, branch lengths, etc.) have been estimated from the data.
2. Experiments with PASSML

PASSML was downloaded from [http://www.ebi.ac.uk/goldman/hmm/passml.html](http://www.ebi.ac.uk/goldman/hmm/passml.html). PASSML uses an HMM approach to model the series of structural and exposure states along the protein sequences. To properly use PASSML, we thus needed rough alignments with all original sites and no removal of gapped regions. For each of our 300 testing alignments, we kept the longest continuous block produced by Gblock when cleaning HSSP (Le et al., Bioinformatics 2008). We then removed all alignments where the number of left sites was less than twice the number of taxa. We obtained 148 alignments with 27 taxa and 147 sites on average. As PASSML is not able to infer tree topologies, we used both LG-Γ4 and LG+Γ4 topologies. Using these topologies, we ran our EX_EHO_MIX and EX_EHO_CONF/MIX models. Results are reported below. All models are compared to LG+Γ4 using the average AIC per site criterion. We see that:

- **LG-Γ4** is by far the worse model, which clearly indicates that a gamma distribution must be used.
- **PASSML** (not using any gamma distribution) improves over LG-Γ4, which shows the strength of its multi-matrix HMM model. However, PASSML is worse than JTT+Γ4.
- **EX_EHO_MIX-Γ4**, which is our model closest from PASSML (no gamma distribution, using structure and exposure categories, but ignoring available annotations), improves over PASSML (likely due to better estimated replacement rate matrices and the fact that category proportions are fitted with each dataset separately), but is similar to JTT+Γ4 and worse than LG+Γ4.
- **EX_EHO_CONF/MIX-Γ4** is slightly better than LG+Γ4, which demonstrates the advantage of using available structural annotations (compared to PASSML and EX_EHO_MIX-Γ4), while **EX_EHO_CONF/MIX+Γ4** is clearly best, as it uses both the structural annotations and gamma distributed rates across sites.

<table>
<thead>
<tr>
<th>Model</th>
<th>AIC/site compared to LG+Γ4</th>
</tr>
</thead>
<tbody>
<tr>
<td>LG-Γ4</td>
<td>-1.802</td>
</tr>
<tr>
<td>JTT+Γ4</td>
<td>-0.333</td>
</tr>
<tr>
<td>PASSML, LG-Γ4 topology</td>
<td>-0.855</td>
</tr>
<tr>
<td>PASSML, LG+Γ4 topology</td>
<td>-0.858</td>
</tr>
<tr>
<td>EX_EHO_MIX-Γ4, LG-Γ4 topology</td>
<td>-0.264</td>
</tr>
<tr>
<td>EX_EHO_CONF/MIX-Γ4, LG-Γ4 topology</td>
<td>0.102</td>
</tr>
<tr>
<td>EX_EHO_CONF/MIX+Γ4, LG+Γ4 topology</td>
<td>0.824</td>
</tr>
</tbody>
</table>
3. Comparison of CONF/LG, PART and MIX

CONF/LG, defined by Equation (3), combines the partition model (PART) and a single-matrix model (LG). Thus, the likelihood values obtained with CONF/LG are necessarily better than those provided by PART and LG separately. The difference in AIC value between CONF/LG and LG is significant for most of our HSSP test alignments (Figure 2). Here, we measure the gain of CONF/LG compared to PART and MIX (mixture model). Figure A shows the number of alignments (among 300 HSSP test alignments) where CONF/LG has significantly better AIC value than PART (KH test on AIC values with p<0.01). Figure B shows the number of alignments where PART and CONF/LG are (significantly, using KH test on AIC values with p<0.01) better/worse than MIX. We see that:

- Despite that CONF/LG clearly improves PART on average (see Figure 1), the difference is rarely significant for any given alignment (Fig. A), specially with EHO (secondary structure) annotations.
- Overall (Fig. B), PART is better than MIX, except with EHO. This confirms the average results of Figure 1.
- In the comparisons with MIX (Fig. B), CONF/LG has a clear effect on PART, decreasing the number of alignments where MIX is best, and augmenting the number of significant positive differences. Again, EHO is the exception, which contrasts with the results of Figure 3 where both PART and MIX are clearly improved by CONF/MIX. This indicates that EHO annotations are relatively poor, but that EHO is a relevant site partition in a phylogenetic perspective.

![Bar chart](image)

**Figure A:** number of alignments where CONF/LG is significantly better than PART.
Figure B: number of alignments where CONF/LG and PART are (significantly) better (positive side) and worse (negative side) than MIX.