

Supplementary Material

1 KM Equations

Following are the KM equations for the Transition matrix elements ($\alpha$, $\beta$, $\gamma$, $\delta$, $\epsilon$) (Materials and Methods section).

$$
\alpha_\eta = \frac{1}{2} p_1^{(\eta)} \left[ 1 - \frac{1}{2} p_2^{(\eta)} \left( \frac{1}{2} - p_4^{(\eta)} \right) \right]
$$

$$
\beta_\eta = 1 - p_1^{(\eta)} \left( 1 - \frac{1}{4} p_2^{(\eta)} p_3^{(\eta)} \right)
$$

$$
\gamma_\eta = \frac{\frac{1}{2} a_\eta p_2^{(\eta)} p_3^{(\eta)} - \left( \frac{7}{8} p_1^{(\eta)} - 1 \right)(1 - a_\eta)}{1 + \frac{1}{2} a_\eta p_2^{(\eta)} \left( \frac{1}{1 - a_\eta} \right)}
$$

$$
\delta_\eta = \frac{\frac{1}{2} (1 - a_\eta) a_\eta p_2^{(\eta)} p_4^{(\eta)} + \frac{3}{8} p_1^{(\eta)} (1 - a_\eta)^2}{1 - a_\eta + \frac{1}{2} a_\eta p_2^{(\eta)}}
$$

$$
\epsilon_\eta = \frac{\frac{1}{2} a_\eta p_2^{(\eta)} \left( p_4^{(\eta)} + \frac{a_\eta}{1 - a_\eta} \right) + \frac{1}{2} p_1^{(\eta)} (1 - a_\eta) + a_\eta}{1 + \frac{1}{2} a_\eta p_2^{(\eta)} \left( \frac{1}{1 - a_\eta} \right)}
$$

The $p$ probabilities are defined as follows:
\[ p_1^{(\eta)} = 1 - e^{-2r_\eta t_\eta}, \]
\[ p_2^{(\eta)} = 1 - p_1^{(\eta)}/2r_\eta t_\eta, \]
\[ p_3^{(\eta)} = (1 - a_\eta)/(1 + a_\eta), \]
\[ p_4^{(\eta)} = a_\eta/(1 + a_\eta), \]

where \( \eta \in \{1, 2\} \).

Probabilities \( p_1^{(\eta)} \) and \( p_2^{(\eta)} \) model the indel rate \( r \), and probabilities \( p_3^{(\eta)} \) and \( p_4^{(\eta)} \) model the indel length distribution \( a \), in region \( \eta \), as in Knudsen and Miyamoto (2003). These four probabilities are compatible with the empirical findings of Pascarella and Argos (1992). They play a role in our modelling by providing coupling between the evolutionary time parameter \( t \) – measured in units of expected substitutions (or replacements) per site – with the indel parameters \( a \) and \( r \) when estimating the vectors \((\alpha_\eta, \beta_\eta, \delta_\eta, \epsilon_\eta, \gamma_\eta), \eta \in \{1, 2\} \), in the two-region model.

2 Data Sets

2.1 Protein Data Set

Figure 1: Boxplot of 808 PIPs extracted from BAliBASE. PIPs have a large fourth spread and a large number of outliers. Lower and upper fourths indicate a suitable data set lying between 0.25 and 1.25.

Figure 1 shows the boxplot of evolutionary distances of the 808 protein PIPs described in the Materials and Methods section. It reveals a spread resistant to outliers ranging from 0.03 to 2.36. To eliminate outliers, we selected a sub-sample of PIPs that had distances within this range, and discarded the rest. In order to reduce the skew in the data, we also subdivided this sub-sample into nine bins, with each bin having a bandwidth of 0.25 (table 1). Bins 2 to 5 were found to have PIPs
Table 1: — The BAliBASE sample of PIPs in bins 1 to 9

| Bin | Size | Min  | Max  | Median | Mean  | Diff-M\(^a\) | SD   | Diff-SD\(^b\) |
|-----|------|------|------|--------|-------|--------------|------|-------------|
| 1   | 99   | 0.027| 0.249| 0.196  | 0.176 | 0.061        |      |             |
| 2   | 185  | 0.250| 0.498| 0.378  | 0.372 | 0.196        | 0.075| 0.014       |
| 3   | 140  | 0.502| 0.749| 0.607  | 0.618 | 0.246        | 0.071| -0.004      |
| 4   | 126  | 0.750| 0.995| 0.875  | 0.871 | 0.254        | 0.068| -0.002      |
| 5   | 78   | 1.004| 1.249| 1.109  | 1.124 | 0.253        | 0.071| 0.003       |
| 6   | 64   | 1.251| 1.497| 1.374  | 1.373 | 0.248        | 0.077| 0.006       |
| 7   | 40   | 1.500| 1.734| 1.611  | 1.617 | 0.244        | 0.072| -0.005      |
| 8   | 25   | 1.761| 1.993| 1.839  | 1.869 | 0.252        | 0.077| 0.005       |
| 9   | 14   | 2.009| 2.153| 2.071  | 2.078 | 0.209        | 0.047| -0.030      |

\(^a\) Difference between the means of two consecutive bins.
\(^b\) Difference between the standard deviations (SD) of two consecutive bins.

whose statistical properties are better; that is, (i) differences in the means are precisely 0.25 and (ii) differences in the standard deviations are uniform and small. At the same time, each of these four bins turned out to have sizes large enough to allow random sampling of 30 PIPs per bin without outliers. Hence we only used these four bins for sampling, and generated pooled experimental samples of 120 protein PIPs with evolutionary distances ranging from 0.25 to 1.25. Boxplots (not shown) typically had lower and upper fourths of 0.5 and 1.0, respectively, and a median typically of 0.75, thus showing that our sampling procedure produced experimental samples that had the uniformity required for statistical inference.

2.2 rRNA Data Set

Figure 2 shows boxplots (A) to (D) of evolutionary distances in the rRNA data. Parts (A) and (B) belong to PIPs extracted from the first and second trees, respectively, as explained in the Materials and Methods section. These boxplots reveal that the spread of evolutionary distances obtained from the two respective trees range between 0.00 and 0.25 after discarding outliers.

In order to double this range, we pruned the second tree several times. That is, first we performed a post-order traversal to obtain PIPs. Second, we pruned leaves corresponding to these PIPs from the group of leaves belonging to the tree. And third, we re-constructed a new tree from the remaining sequences in the group. We repeated these three steps until the tree was reduced to a size whereby the next PIP yielded a distance which exceeded a set upper limit of 0.50. We chose this upper limit because it provided a sufficient number of PIPs with a median distance close to 0.25.

This procedure yielded 105 PIPs with unique sequences, and their evolutionary distances are plotted in part (C). The combined boxplot of parts (A) and (C) is shown in part (D). The combined boxplot reveals that the spread was somewhat negatively skewed, with a median of 0.21 and no outliers. To compensate for the skew, we retained all the 74 PIPs from the first tree (boxplot (A))
Figure 2: Boxplots of rRNA PIPs. Parts (A) and (B) belong to the 74 and 151 unique PIPs, respectively. Part (C) belongs to the 105 unique PIPs obtained after pruning. Part (D) belongs to the combined spread of PIPs in parts (A) and (C).

and added a random sample of 25 PIPs from among the 105 PIPs that had been obtained from the pruning procedure (boxplot (C)). This produced an experimental sample of 99 rRNA PIPs.

3 Model Validation

Each parameter in the two-region model (Materials and Methods section) was tested for statistical power. To carry out the test, sets of 12 pairwise alignments (as described in the Results section) were simulated under a regime of arbitrary parameter values shown in table 2. In this table, each of the three row cells corresponding to (i) parameter value settings in column 2 and to (ii) column headings 3, 4, and 5 shows a p-value obtained from a large-sample t test defined as $H_0: \mu - \mu_o = 0$ versus $H_a: \mu - \mu_o \neq 0$ (where $\mu_o$ is the sample mean). Each of these p-values was computed from a data set consisting of 24 point estimators; that is, two ML estimators per each simulated pair in the corresponding set. Estimation was carried out by optimising the likelihood function for each simulated pair over the corresponding two parameters (one in each region), while keeping all other
parameters fixed at their true value.

Table 2: — Large-sample t tests (Protein Alphabet)

| Parameters                        | Settings  | $t_2 = 0.4$ | $t_2 = 0.6$ | $t_2 = 0.8$ |
|-----------------------------------|-----------|-------------|-------------|-------------|
| replacement rate ($t_1$)          | (0.20)    | 0.1309      | 0.1865      | 0.5543      |
| indel length distribution ($a_1, a_2$) | (0.20, 0.30) | 0.5462      | 0.2572      | 0.3318      |
| indel rate ($r_1, r_2$)           | (0.30, 0.05) | 0.1585      | 0.6143      | 0.8128      |
| hydrophilicity ($h_1, h_2$)       | (0.40, 0.70) | 0.5202      | 0.0488      | 0.7110      |

Note 1. To simulate pairwise alignments, $t_1$ was arbitrarily set to 0.2, and $t_2$ was set to arbitrary values 0.4, 0.6, and 0.8 as shown in column headings 3, 4, and 5 respectively. For each combination of $t_1$ and $t_2$, a set of 12 alignments was simulated, while all other parameters shown in column 1 were set to corresponding arbitrary values shown in column 2.

Note 2. In row 1, $t_1$ and $t_2$ were allowed to vary from their true value to compute $p$-values in columns 3 - 5, while all other parameters were held constant.

Note 3. Similarly, in row 2, $a_1$ and $a_2$ were allowed to vary, and so on.

Note 4. Each $p$-value in columns 3 - 5 was computed using a large-sample t test with 24 data points as explained in the text.

Results obtained from amino acid alignments of simulated pairs are shown in table 2. Similar results (not shown) were also obtained from DNA alignments after replacing the PMB model with the HKY model. In both cases, the large-sample $t$ tests showed that all expected values were not significantly different from the corresponding true values at the 5% level of significance.

Table 3: — ANOVA (Protein Alphabet)

| Source of Variation | d.f. | SS | MS | f-test | p-value |
|---------------------|------|----|----|--------|---------|
| Parameters          | 3    | 2.42 | 0.81 | 1.1969  | 0.3113  |
| Distances           | 2    | 0.69 | 0.35 | 0.5119  | 0.5999  |
| Interaction         | 6    | 2.71 | 0.45 | 0.6701  | 0.6739  |
| Error               | 276  | 186.29 | 0.68 |
| Total               | 287  | 192.12 |

(a) Source of variation
(b) Degrees of freedom
(c) Sum of squares
(d) Mean square
(e) Interaction is between parameters and distances, and has $(4 - 1) \times (3 - 1) = 6$ d.f.

A two-factor analysis of variance (ANOVA) was also carried out on the two-region model to test (i) for main effects from different parameters (namely, $t$, $a$, $r$, and $h$), (ii) for main effects from different nominal distances (namely, 0.4, 0.6, and 0.8), and (iii) for interaction between these two effects. The data matrix (not shown) for the ANOVA consisted of four rows (one for each parameter) and three columns (one for each nominal distance). Each cell in the data matrix was the average of the 24 ML point estimators obtained from the corresponding simulated pairs.

The $p$-value of 0.67 shown in row three of the last column of ANOVA table 3 shows that the two-region model has no significant interaction between the two types of effects. Main effects from
parameters in the model were not significant, with a $p$-value of 0.31 shown in row one. Similarly, main effects from different nominal distances were also not significant, with a $p$-value of 0.60 shown in row two. These results show that the two-region model is not adversely affected by different parameters varying separately and independently in the two regions over a wide range of different nominal distances using the PMB evolutionary model.

| SOV         | d.f. | SS     | MS   | $f$-test | $p$-value |
|-------------|------|--------|------|----------|-----------|
| Parameters  | 3    | 93.62  | 31.21| 3.3933   | 0.0184    |
| Distances   | 2    | 33.41  | 16.71| 1.8165   | 0.1645    |
| Interaction | 6    | 68.69  | 11.45| 1.2448   | 0.2835    |
| Error       | 276  | 2538.18| 9.20 |          |           |
| Total       | 287  | 2733.89|      |          |           |

A similar ANOVA was carried out using the DNA simulated data and replacing the PMB model with the HKY model. The four parameters in this case were $t$, $a$, $r$, and $\kappa$. $p$-values in ANOVA table 4 show no significant interaction between parameter effects and distance effects, and no significant main effects from different nominal distances. However, from the $p$-value of 0.0184 in row one, main effects from different parameters were not significant at the 1% level of significance, but were significant at the 5% level of significance. We attribute the presence of these parameter effects to the transition-transversion rate ratio parameter $\kappa$ and long sequences in the simulated data. We do not consider this result to be a major weakness of the two-region model. This is because we applied the two-region model to rRNA data having a range of evolutionary distances much narrower than that of our amino acid data.