BAYESIAN COMPARISONS OF CODON SUBSTITUTION MODELS

SUPPLEMENTARY MATERIAL

Nicolas Rodrigue, Nicolas Lartillot and Hervé Philippe
Thermodynamic integrations

We used the model-switch thermodynamic integration framework proposed in Lartillot and Philippe (2006) to evaluate Bayes factors across all codon substitution models described in the main text. We summarize the general framework here, and describe the specific instances used for the models under study.

General model-switch

Given a data set \( D \) and a model \( M \), specified by some high-dimensional parameter vector \( \theta \in \Theta \), a Bayesian analysis begins with an evaluation of the posterior distribution \( p(\theta \mid D, M) \), obtained from the likelihood function \( p(D \mid \theta, M) \) and the prior \( p(\theta \mid M) \) according to Bayes’ theorem:

\[
p(\theta \mid D, M) = \frac{p(D \mid \theta, M)p(\theta \mid M)}{p(D \mid M)}, \tag{1}
\]

where

\[
p(D \mid M) = \int_{\Theta} p(D \mid \theta, M)p(\theta \mid M)d\theta \tag{2}
\]

is the marginal likelihood, sometimes also called the prior predictive probability. For all models of interest here, the integral in (2) has no analytical form. However, Markov chain Monte Carlo approaches allow one to sample from the posterior distribution of parameters of interest, without knowing the marginal likelihood, which cancels out of the basic Metropolis-Hastings kernel (Metropolis et al., 1953; Hastings, 1970): given the current parameter configuration \( \theta \), generate a
new parameter configuration \( \theta^\prime \) from the density \( q(\theta, \theta^\prime) \), and set \( \theta = \theta^\prime \) with probability \( \vartheta \):

\[
\vartheta = \min\left\{ 1, \frac{p(\theta^\prime \mid D, M)}{p(\theta \mid D, M)} \frac{q(\theta^\prime, \theta)}{q(\theta, \theta^\prime)} \right\}.
\] (3)

Following a “burn-in” period, the Markov chain formed by cycling over this procedure is used to produce a large sample approximation of the posterior distribution. All likelihood computations and MCMC update operators necessary for implementing the models studied in this work have been described previously (e.g., Larget and Simon, 1999; Felsenstein, 2004; Huelsenbeck and Dyer, 2004; Huelsenbeck et al., 2006; Yang, 2006).

Our objective here is to compare two models, \( M_0 \) and \( M_1 \), based on the Bayes factor \( (B_{01}) \), defined as the ratio of their respective marginal likelihoods:

\[
B_{01} = \frac{p(D \mid M_1)}{p(D \mid M_0)}
\] (4)

A Bayes factor greater than (less than) 1 is considered as evidence in favor of \( M_1 (M_0) \). However, because the basic MCMC algorithms described above are explicitly designed to avoid computing marginal likelihoods, more elaborate methods are needed.

The model-switch thermodynamic integration method extends the advantages of MCMC sampling by devising a path linking the posterior distributions of two models. Let \( \theta \) now represent the union of parameters from both models (some of which may be relevant to both models, e.g., branch lengths, nucleotide exchangeabilities, while others are only relevant to one of the two models). Two
models of interest can be connected by defining

\[ p(D \mid \theta, M_\beta) = e^{(1-\beta) \ln p(D \mid \theta, M_0) + \beta \ln p(D \mid \theta, M_1)}, \] (5)

\[ p(\theta \mid M_\beta) = e^{(1-\beta) \ln p(\theta \mid M_0) + \beta \ln p(\theta \mid M_1)}, \] (6)

\[ p(\theta \mid D, M_\beta) = \frac{p(D \mid \theta, M_\beta) p(\theta \mid M_\beta)}{p(D \mid M_\beta)}, \] (7)

and the Metropolis-Hastings kernel as

\[ \vartheta = \min \left\{ 1, \frac{p(\theta' \mid D, M_\beta) q(\theta', \theta)}{p(\theta \mid D, M_\beta) q(\theta, \theta')} \right\}. \] (8)

For any value \(0 < \beta < 1\), the kernel given in (8) allows one to sample from a posterior distribution consisting of a partial “morphing” between \(M_0\) and \(M_1\), without knowing \(p(D \mid M_\beta)\). The quasi-static method described in Lartillot and Philippe (2006) initially sets to \(\beta = 0\), and the resulting sampler has the posterior of parameters under \(M_0\) as its limiting distribution. Then, the value of \(\beta\) is regularly incremented by a small value \(\delta\beta\) after a set of MCMC cycles, until \(\beta = 1\); the sampler finally has the posterior under \(M_1\) as its limiting distribution. Note that here, we do not explore models with different priors on the same parameters, and hence we can dispense with the morphing prior defined in (6), substituting it with \(p(\theta \mid M_0, M_1)\). When calling Metropolis-Hastings operators on components of \(\theta\) that are only relevant to \(M_0\), the prior can be reduced to \(p(\theta \mid M_0, M_1) = p(\theta \mid M_0)\); and likewise when calling operators on components relevant only to
$M_1$, in which case $p(\theta \mid M_0, M_1) = p(\theta \mid M_1)$. Based on a sample collected along the entire path of posterior distributions, written as $(\beta_h, \theta_h)_{h=0..K}$, where $\beta_0 = 0$, $\beta_K = 1$ and $\forall h, 0 \leq h < K$, $\beta_{h+1} - \beta_h = \delta \beta$, the (log) Bayes factor between $M_0$ and $M_1$ can be estimated based on the Monte Carlo relation:

$$
\ln B_{01} = \ln p(D \mid M_1) - \ln p(D \mid M_0) \\
= \int_0^1 \langle \ln p(D \mid \theta, M_1) - \ln p(D \mid \theta, M_0) \rangle_{\beta} d\beta \\
\simeq \frac{1}{K} \left[ \frac{1}{2} \left( \ln p(D \mid \theta_0, M_1) - \ln p(D \mid \theta_0, M_0) \right) + \right. \\
\left. \left( \sum_{h=1}^{K-1} \ln p(D \mid \theta_h, M_1) - \ln p(D \mid \theta_h, M_0) \right) + \right. \\
\left. \frac{1}{2} \left( \ln p(D \mid \theta_K, M_1) - \ln p(D \mid \theta_K, M_0) \right) \right],
$$

where $\langle \cdot \rangle_{\beta}$ stands for an expectation with respect to (7).

The overall precision of the method depends on a number of factors, such as the step size ($\delta \beta$), and whether the number of cycles between steps is sufficient to allow the chain to re-equilibrate to (7), for instance; but also on the inherent distance between the two models being compared. With a large set of candidate models, a reasonable traversal across the space of all models must be designed for efficient computation. In the following sub-sections, we describe a set of of model-switch thermodynamic integrations linking together all models under study.

**GY-MG-switch**

The first model-switch scheme links together the GY-F1×4 and the MG-F1×4 models. This particular thermodynamic integration represents the ideal case, where all parameters are involved in
both models; parameters are always sampled from the posterior distribution of one model or the other (or the partially morphed posteriors along the path). The GY-MG-switch is also applied to link GY-F3×4 and MG-F3×4 models.

**F1×4-F3×4-switch**

The F1×4-F3×4-switch is only used in the GY context, although it could be used in the MG context as well; here, only one of the two contexts need be calculated to link all models together. For this model-switch, the single nucleotide frequency vector of the GY-F1×4 model is also used as the first codon position nucleotide vector under the GY-F3×4 model. As such, at one end of the path, this set of nucleotide frequencies corresponds to the single-nucleotide-vector-approximation of codon frequencies under the GY-F1×4 model, whereas at the other end, it corresponds to the first position vector of the three-vector-approximation of codon frequencies under the GY-F3×4 model. As for the other two nucleotide vectors associated with the the GY-F3×4 model, they are effectively sampled from the prior at one end of the path, and the posterior at the other end of the path. All other parameters are relevant to both models.

**F1×4-F61-switch**

This model-switch is only pertinent to the GY context, connecting the GY-F1×4 model and the the GY-F61 model. At one end of the path, the vector of nucleotide propensities used to approximate codon frequencies is sampled from the posterior under the GY-F1×4 model, while sampling from the prior of a (distinct) full 61-dimensional codon frequency vector. At the other end of the path, the vector of nucleotide frequencies used to approximate codon frequencies is sampled from the
prior, whereas the 61-dimensional codon frequency vector is sampled from the posterior. All other parameters are relevant to both models.

**CP-switch and AAP-switch**

The CP-switch is only pertinent in the MG context, linking the MG-F1×4 and MG-F1×4-CP models. One end of the path samples the codon preference parameters from the prior, whereas the other samples these parameters from the posterior. All other parameters are relevant to both models. The CP-switch is also used to link the MG-F3×4 and MG-F3×4-CP models. The AAP-switch is analogous to the CP-switch, but involving the amino acid preference parameters instead.

**DP-switch**

This last model-switch links together a model with a single $\omega$ factor and a model based on the Dirichlet process prior modeling heterogeneous $\omega$ factors across sites. At one end of the path, the sampler draws from the posterior of a model with a single $\omega$ factor, and from the prior (and hyper-prior) of the Dirichlet process. At the other end of the path, sampling is under the full posterior of the Dirichlet process, and the prior of the global $\omega$ factor. As before, all other parameters are relevant to both models. The DP-switch scheme is applied separately to each underlying GY and MG-style model.
Overall model ranking

From the set of model-switch methods described above, we can evaluate all models by computing Bayes factors with respect to a common reference. We use GY-F1×4 as the reference model here, which implies that as many as 4 different sets of model-switch schemes may be involved in reporting a particular Bayes factor. For instance, taking the example from the main text, the (log) Bayes factor between MG-F3×4-CP-DP and GY-F1×4 is assembled from four separate calculations:

\[
\ln \frac{p(D \mid \text{MG-F3×4-CP-DP})}{p(D \mid \text{GY-F1×4})} = \ln \frac{p(D \mid \text{MG-F3×4-CP-DP})}{p(D \mid \text{MG-F3×4-CP})} + \ln \frac{p(D \mid \text{MG-F3×4-CP})}{p(D \mid \text{MG-F3×4})} + \ln \frac{p(D \mid \text{MG-F3×4})}{p(D \mid \text{GY-F3×4})} + \ln \frac{p(D \mid \text{GY-F3×4})}{p(D \mid \text{GY-F1×4})},
\]

where the first term is computed using the DP-switch, the second using the CP-switch, the third using the GY-MG-switch, and the fourth using the F1×4-F3×4-switch.

Direct-switch

As an alternative to performing the multi-step model space traversal described above, we could run a single (bidirectional) model-switch from each model directly to the reference model. As a specific example, we tried a model switch linking the highest dimensional model (MG-F3×4-CP-DP) to the reference model (GY-F1×4): at one end of the path, the three vectors of nucleotide propensities, the codon preference parameters, and the Dirichlet process on nonsynonymous rate factors are sampled from the prior, with the remaining parameters sampled from the posterior under the GY-F1×4.
The other end of path samples from the posterior under the MG-F3×4-CP-DP model, and from the prior of the vector of nucleotide propensities used to approximate codon stationary probabilities and the global nonsynonymous rate factor. All other parameters are relevant to both models.

**Empirical explorations of thermodynamic integrations**

We performed several pilot runs to tune each type of model-switch thermodynamic integration. Incorporating the bi-directional approach described in Lartillot and Philippe (2006), each model-switch scheme was explored by running integrations in duplicates, one with $\beta$ going from 0 to 1, and another with $\beta$ going from 1 to 0. We report both values obtained from the bi-directional approach as an interval throughout. The idea here is that (with the exception of the GY-MG-switch) at the beginning of the forward run ($\beta$ going from 0 to 1), the set of parameters relevant to one of the models ($M_1$) is being sampled from the prior distribution, and as the run progresses, the sampler “anneals” into the posterior of the these parameters. Conversely for the parameters relevant to the other model ($M_0$), the sampling is progressively “melted”, from draws under the posterior to draws under the prior. If the sampler is not allowed to properly equilibrate between each increment of the run, the parameters relevant to $M_1$ will retain aspects of the sampling under the prior (most of which will have very weak log-likelihoods), and thus tend to over-estimate the penalty of using such a set of parameters. Conversely for the parameters of $M_0$, at the end of the run, these will have retained aspect of the sampling under the posterior (most of which have rather good log-likelihoods), and hence tend to under-estimate the penalty of these parameters. Altogether, the forward run will tend to under-estimate the Bayes factor in favor of $M_1$. The backward run ($\beta$
going from 1 to 0) exhibits the exact reverse situation, and tends to over-estimate the Bayes factor in favor of $M_1$.

Figures S1 and S2 display examples of this tuning process in two cases. Each panel in these figures plots the values $\ln p(D | \theta, M_1) - \ln p(D | \theta, M_0)$ collected during bi-directional quasi-static runs. Graphically, the log Bayes factor corresponds to the area between the curve and the abscissa (negative below the abscissa, and positive above it), and is estimated using the relation given in (11).

**An easy example**

Using the $\beta$-Globin17-144 data set, figure S1 corresponds to a case that we qualify as computationally easy: the GY-MG-switch, linking GY-F1×4 and MG-F1×4. These two models have the exact same parameters, and only differ in how parameters are assembled to specify the final model. At one end of the path ($\beta \sim 0$), the plot displays the difference in log-likelihood between MG-F1×4 and GY-F1×4, when the parameters from the posterior under GY-F1×4 are “imposed upon” the MG-F1×4 model. Reciprocally, at the other end of the path ($\beta \sim 1$), the plot displays the difference in log-likelihood when the parameters of the posterior under MG-F1×4 are “imposed upon” GY-F1×4 model. In other words, for the model-switch, no parameters are ever drawn from the prior (at either end of the path). Based on the $K+1$ draws along the path, the approximation given in (11) for $K = 100$ (fig. S1a), $K = 1,000$ (fig. S1b), and $K = 10,000$ (fig. S1c), is $[2.9 ; 5.4]$, $[3.7 ; 4.1]$ and $[3.8 ; 3.9]$ respectively. These two models are quite close to each other, in terms of overall fit, but the model-switch integration procedure nonetheless allows for a very precise estimation in this case, because the models can be connected through a very short overall path. In
this case, the final runs ($K = 10,000$) each required about 6 days of CPU time on an Intel P4 3.2 GHz computer node.

**A challenging example**

Still using the $\beta$-GLOBIN17-144 data set, figure S2 corresponds to a case that we qualify as computationally challenging: the F1×4-F61-switch, linking GY-F1×4 and GY-F61. In contrast with the GY-MG-switch, in which all parameters were involved in both models, this thermodynamic integration has a set of parameters in each model that are irrelevant to the other. When $\beta \approx 0$, the plots display the difference in log-likelihood between GY-F61 and GY-F1×4 when the 61-dimensional vector of codon frequencies attributed to GY-F61 is sampled from the prior, and other parameters are those “imposed by” the posterior under GY-F1×4. Such a sampler will induce very poor log-likelihood values under GY-F61, and indeed the plots display negative values at this end of the path. At the other end of the path ($\beta \approx 1$), the plots display the difference in log-likelihood between GY-F61 and GY-F1×4 when the 61-dimensional vector of codon frequencies is sampled from the posterior under GY-F61, the single-nucleotide-vector-approximation of codon frequencies is sampled from the prior under GY-F1×4, but with other parameters being those “imposed” by the posterior under GY-F61. Thus, this other end of the path will induce very poor log-likelihood values under GY-F1×4, and indeed the log-likelihood difference displayed in the plots become highly positive. As can be appreciated graphically, the tail-ends of the integrand represent the main source of error in this model switch. The interval obtained from bi-directional quasi-static runs with $K = 100$ (fig. S2a) is extremely broad, at [84.9 ; 175.6]. Several tuning options could be explored, but here, we simply increase the overall sample size (or equivalently, decrease the step size
\(\delta \beta\). With \(K = 1,000\) (fig. S2b) the interval obtained is \([113.4 ; 126.2]\), and finally, the longest runs (\(K = 10,000\), fig. S2c), each requiring about 20 days of CPU time, produce the tightest interval, at \([115.6 ; 117.6]\).

**The Direct-switch**

We performed the same empirical tuning for the direct model switch, linking the highest dimensional model (MG-F3×4-CP-DP) to the reference model (GY-F1×4). Using the \(\beta\)-GLOBIN 17-144 data set, figure S3 plots the recorded log-likelihood differences over three different bidirectional runs. With \(K = 100\) (fig. S3a), the interval from bidirectional calculations is \([49.4 ; 521.1]\). With \(K = 1,000\) (fig. S3b) the interval is \([196.4 ; 252.9]\). Finally, with \(K = 10,000\) (fig. S3c), the interval is \([237.8 ; 242.3]\), which corroborates (and is overlapping with) the result obtained using the multi-step calculation. However, these last two direct thermodynamic samples, each required about 53 days of CPU. Altogether, the multi-step calculation of this log Bayes factor required about 49 days, with a similar level of precision, but also computed three other log Bayes factors along the way. For these reasons, we did not further pursue the direct model-switch scheme, although we note that it may be pertinent in other contexts, depending on the set of models of interest.

**Further comparisons of posterior distributions**

Figure S4 explores the posterior distribution of the F3×4 configuration under the MG-F3×4-DP and MG-F3×4-CP-DP models. First note that without the CP parameters (full lines), the three positions show striking differences in overall distributions, and that the magnitude of the credibility
intervals are much greater than under the F1×4 configuration. When the CP parameters are introduced (dashed lines), several credibility intervals considerably shift and increase in magnitude. Also note that under the CP settings, the distributions of each position tend to overlap. To show this more vividly, we reconstituted the 95% credibility intervals of global and position-specific nucleotide propensity parameters into a single figure for the Globin17-144 data. Figure S5a displays the global nucleotide propensity parameter values obtained under the the MG-F1×4-DP model (full line) as well as each of the three nucleotide propensity parameter values under the MG-F3×4-DP (in progressively finer dashed lines for position 1, 2, and 3). In this case, the disparity between the different distributions is high. When including the CP parameters (fig. S5b), however, the disparity is much lower, suggesting the redundancy of the F3×4 configuration when invoking the CP parameters. We note that some values are still markedly divergent (e.g., 3rd position A and 2nd position T), indicating that other model violations may be at play. In other words, codon (or amino acid) preferences may explain (albeit perhaps not entirely) the observed disparities of nucleotide frequencies at the three codon positions.

Figure S6 compares the 95% credibility intervals of codon preference parameters under the MG-F1×4-CP-DP (full line) and MG-F3×4-CP-DP (dashed line) models. The parameters appear to be only moderately sensitive to the F1×4/F3×4 choice, although a few notable shifts and increases in magnitude of credibility interval are observed, e.g., CTG, CTC, and GTG; these three examples in particular may be suggestive of identifiability problems, specifically with the F3×4 configuration, which indeed leads the distribution of the second position T propensity parameter to drop (fig. S5b). Nonetheless, the general stability of the CP distributions to the F1×4/F3×4 choice suggests that the although the F3×4 configuration is impertinent with the CP parameters, it is not too
costly in terms over-parameterization, which corroborates with the computed log Bayes factors.

**POSTERIOR PREDICTIVE CHECKS ON NUCLEOTIDE FREQUENCIES**

Figure S7 displays the posterior predictive distributions of nucleotide frequencies obtained under the MG-F1×4-DP (full-line histograms) and MG-F1×4-CP-DP (dashed-line histograms) models, for the GLOBIN\(17-144\) data set. In figures S7a to S7d, which correspond to the global nucleotide frequency posterior predictive distributions, we first note that for the MG-F1×4-DP model we obtain a reasonable matching with the observed values, although the \(p\)-value does approach 1 in panel S7b. This provides a first (albeit weak) indication that utilizing a nucleotide propensity vector on its own is problematic, and suggests that other unaccounted features may be at play. In contrast, we note that the MG-F1×4-CP-DP model seems to match well across all dimensions.

When considering codon-position-specific nucleotide frequencies (fig. S7e to p), the MG-F1×4-DP model leads to relatively even frequencies at the three positions (although, as mentioned in the main text, the exclusion of stop codons induces slight shifts), and in many cases, the distributions show a pronounced discrepancy with the empirical values. For the MG-F1×4-DP model values near 0 or 1 are obtained in panels S7f, g, h, i, j, k, m, and o. In contrast, the MG-F1×4-CP-DP model leads to markedly uneven nucleotide frequencies at each of the three positions, with most distributions encompassing the codon-position-specific empirical values. A few exceptions, (fig. S7m and o), are indicative of remaining problems with the model.

In figure S8, we repeated this posterior predictive experiment with the GY-F61-DP model (full-line histograms), contrasting the results with those obtained under the MG-F1×4-CP-DP model
(dashed-line histograms). For the global nucleotide frequencies (fig. S8a to d), we find that the GY-F61-DP model induces distributions that reasonably encompass the empirical values in two cases (fig. S8a and b), whereas the distributions are slightly deviated in two others (fig. S8c and d). With regards to codon-position-specific nucleotide frequencies, we note that the GY-F61-DP model indeed induces uneven propensities at each position (as we noted in the main text), but some panels indicate that it does not perform as well as the MG-F1×4-CP-DP model (e.g., fig. S8g and p).


Figure S 1. Log-likelihood differences recorded during GY-MG-switch thermodynamic integrations linking GY-F1×4 and MG-F1×4. Two integrations are plotted in each panel, one with $\beta$ going from 0 to 1 (+), and another with $\beta$ going from 1 to 0 ($\times$). The collection of $K+1$ values is used to approximate the log Bayes factor according to (11). Panel a) displays “fast” runs, with $K = 100$, panel b) displays “medium” runs, $K = 1,000$, and panel c) displays “slow” runs, with $K = 10,000$. 
Figure S 2. Log-likelihood differences recorded during F1×4-F61-switch thermodynamic integrations linking GY-F1×4 and GY-F61. Two integrations are plotted in each panel, one with $\beta$ going from 0 to 1 (+), and another with $\beta$ going from 1 to 0 ($\times$). The collection of $K+1$ values is used to approximate the log Bayes factor according to (11). Panel a) displays “fast” runs, with $K = 100$, panel b) displays “medium” runs, $K = 1,000$, and panel c) displays “slow” runs, with $K = 10,000$. 
Figure S 3. Log-likelihood differences recorded during Direct-switch thermodynamic integrations linking GY-F1×4 and MG-F3×4-CP-DP. Two integrations are plotted in each panel, one with $\beta$ going from 0 to 1 (+), and another with $\beta$ going from 1 to 0 (×). The collection of $K + 1$ values is used to approximate the log Bayes factor according to (11). Panel a) displays “fast” runs, with $K = 100$, panel b) displays “medium” runs, $K = 1,000$, and panel c) displays “slow” runs, with $K = 10,000$. 
**Figure S 4.** 95% credibility intervals of position-specific nucleotide propensity parameters obtained under MG-F3\(\times 4\)-DP (full lines) and under MG-F3\(\times 4\)-CP-DP (dashed lines). The three panels (a, b, c) refer to the GLOBIN17-144 data set, followed by LYSIN25-134 (d, e, f), and HIV22-99 (g, h, i).
Figure S 5. A composite a global and position-specific nucleotide propensity parameter for the GLOBIN17-144 data set. Panel a displays the 95% credibility intervals of global nucleotide propensity parameters under the MG-F1×4-DP model (full line) as well as the 95% credibility of the three nucleotide propensity parameters under the MG-F3×4-DP (with progressively finely-dashed lines for position 1, 2, and 3 respectively). Panel b displays the 95% credibility interval for same parameters but now, under the MG-F1×4-CP-DP and MG-F3×4-CP-DP models.
Figure S 6. 95% credibility intervals of codon preference parameters, sorted according to amino acids. The full lines are values under MG-F1×4-CP-DP, whereas the dashed lines are values under MG-F3×4-CP-DP. The leftmost panel (a) refers to the GLOBIN17-144 data set, followed by LYSIN25-134 (b), and HIV22-99 (c).
Figure S 7. Posterior predictive distributions of nucleotide frequencies for the GLOBIN17-144 data set under the MG-F1×4-DP (full-line histograms) and MG-F1×4-CP-DP (dashed-line histogram) models. The top four panels (a, b, c, and d) correspond to the global nucleotide frequencies, followed by the first (e, f, g, and h), second (i, j, k, and l), and third position frequencies (m, n, o, and p). The observed values (computed on the real alignment) are displayed as dashed vertical lines. In the top-right corner of each panel, the first value is the posterior predictive $p$-value under the MG-F1×4-DP model, whereas the second value (below) is that under the MG-F1×4-CP-DP model.
Figure S 8. Posterior predictive distributions of nucleotide frequencies for the GLOBIN 17-144 data set under the GY-F61-DP (full-line histograms) and MG-F1×4-CP-DP (dashed-line histogram) models. The top four panels (a, b, c, and d) correspond to the global nucleotide frequencies, followed by the first (e, f, g, and h), second (i, j, k, and l), and third position frequencies (m, n, o, and p). The observed values (computed on the real alignment) are displayed as dashed vertical lines. In the top-right corner of each panel, the first value is the posterior predictive $p$-value under the GY-F61-DP model, whereas the second value (below) is that under the MG-F1×4-CP-DP model.