Systematic Biology Advance Access published July 29, 2011

Point of View

DOI:10.1093/sysbio/syr065

Fast Bayesian Choice of Phylogenetic Models: Prospecting Data Augmentation–Based Thermodynamic Integration

NICOLAS RODRIGUE[1,2,3,*] AND STÉPHANE ARIS-BROSOU[1,4]

[1]*Department of Biology and Center for Advanced Research in Environmental Genomics, University of Ottawa, 30 Marie Curie Pvt., Ottawa, ON, Canada K1N 6N5;* [2]*Quebec Center for Biodiversity Science, McGill University, 1205 Drive, Penfield Avenue, Montreal, QC, Canada H3A 1B1;* [3]*Eastern Cereal and Oilseeds Research Center, Agriculture and Agri-Food Canada, 960 Carling Avenue, Ottawa, ON, Canada K1A 0C6; and* [4]*Department of Mathematics and Statistics, University of Ottawa, Ottawa, ON, Canada K1N 6N5;*
**Correspondence to be sent to: Eastern Cereal and Oilseeds Research Center, Agriculture and Agri-Foods Canada, 960 Carling Avenue, Ottawa, ON, Canada K1A 0C6; E-mail: nicolas.rodrigue@agr.gc.ca.*

Received 1 November 2010; reviews returned 26 January 2011; accepted 28 February 2011
Associate Editor: Mark Holder

With the increasing number of substitution models being proposed over recent years, there is a need for accurate and fast methods for performing model selection in phylogenetics. However, the most popular computational approach used in Bayesian phylogenetic contexts—the Bayes factor approximated with the harmonic mean estimator (HME) (Newton and Raftery 1994)—has proved unreliable (Lartillot and Philippe 2006; Fan et al. 2011; Xie et al. 2011). Here, we discuss recent advances in reliable computational methods based on thermodynamic integration principles for Bayesian model selection and emphasize the potential of data augmentation–based methods for producing fast and accurate results.

In the Bayesian framework, the evidence in favor of one model over another can be quantified through the Bayes factor, defined as the ratio of two marginal likelihoods (Jeffreys 1935; Kass and Raftery 1995):

$$B_{01} = \frac{p(D|M_1)}{p(D|M_0)}, \tag{1}$$

where D is the available data, and M_0 and M_1 are the two models being compared. A Bayes factor greater than 1 is considered as evidence in favor of M_1. The marginal likelihood under a given model is in fact the expectation of the likelihood with respect to a prior probability distribution:

$$p(D|M) = \int_{\Theta} p(D|\theta, M)p(\theta|M)\mathrm{d}\theta, \tag{2}$$

where θ represents the hypothesis vector, $p(D|\theta, M)$ is the likelihood function, and Θ is the set of all possible hypothesis vectors defined by the prior $p(\theta|M)$.

The marginal likelihood appears as the normalizing constant of the posterior probability distribution in Bayes' theorem:

$$p(\theta|D, M) = \frac{p(D|\theta, M)p(\theta|M)}{p(D|M)}. \tag{3}$$

Although the integral given in Equation (2) often has no analytical solution, Markov chain Monte Carlo (MCMC) approaches have been broadly used to sample from Equation (3) in many contexts, without calculating the marginal likelihood. However, these approaches do not address the issue of quantitative Bayesian model comparison in a practical way. For instance, the most intuitive Monte Carlo approximation that follows from the form of Equation (2), that of producing a sample from the prior $p(\theta|M)$ and taking the average likelihood over such a sample, is not reliable, as high-likelihood regions will tend to be unduly underrepresented in such a sample.

The HME to the marginal likelihood was proposed as a method requiring nothing more than the output of the MCMC system for sampling from the posterior in Equation (3) (Newton and Raftery 1994). The popularity of the HME, as exemplified in the recent phylogenetics literature (e.g., Bleidorn et al. 2007; Brown and Lemmon 2007; Alekseyenko et al. 2008; Praz et al. 2008; Hampl et al. 2009), is undoubtedly due to its methodological simplicity and to its availability in a number of user-friendly programs, such as MrBayes (Ronquist and Huelsenbeck 2003) and BEAST (Drummond and Rambaut 2007). This continued practice is unfortunate, however, given the problems with the HME. Lartillot and Philippe (2006), for instance, convincingly showed that the HME and its so-called stabilized version are unreliable. They proposed the use of another

MCMC-based method, called *thermodynamic integration* (TI), and demonstrated its reliability in analytical contexts. Moreover, Lartillot and Philippe (2006) showed that Bayes factors computed with the HME and TI can produce drastically different results in amino acid substitution model comparisons, to the point of leading to opposite conclusions in model ranking.

Unfortunately, the reliability of TI comes with the cost of greater computing time. Lartillot and Philippe (2006) suggested that TI could require a computational time about one order of magnitude greater than that needed for a plain MCMC run under the most complex of the models being evaluated. It seems likely that TI has not been widely adopted partly due to this increased computing cost. Recent efforts have been made by Xie et al. (2011) and Fan et al. (2011) to speed up TI-based calculations with an approach they call *stepping-stone* (SS). Their basic idea takes advantage of both TI-inspired schemes and importance sampling, in a manner showing great computational promise. In particular, their approach enables reliable quantitative Bayesian model ranking wherever substitution models are tractable in the full pruning-based (also called peeling-based or dynamic programming–based) Bayesian MCMC framework.

The pursuit of general estimation approaches (such as TI and SS) for performing Bayesian model ranking is illustrative of a tendency among MCMC developers and practitioners: that of striving for computational systems that are independent of the particular modeling context. However, with regard to the means of computing Bayes factors, so long as accurate methods are used (thus excluding the HME), generality is not in itself necessary. For now, given the technical challenge of the calculation, it might be warranted to use less general approaches in some modeling contexts, as we illustrate below.

Unreliability of the HME and the Necessity of Alternative Computation

Following Lartillot and Philippe (2006), Fan et al. (2011), and Xie et al. (2011), we demonstrate the issue at stake when using the HME approach to model ranking with a set of codon models described in detail in Appendix 1. The codon substitution models of interest to us here build upon the approach proposed by Muse and Gaut (1994). We write MG to refer to a codon substitution formulation based only on a set of nucleotide-level mutational parameters (see Appendix 1). As presented by Yang and Nielsen (2008), the models of interest are based on the *mutation–selection* principle and parameterize a substitution process in terms of mutational propensities combined with a fixation factor that distinguishes between the types of (nonsynonymous) events. Based on the model nomenclature explained in Appendix 1, MG-MutSelYN is defined as a global codon substitution model, where all model parameters are common to all codon sites in the alignment. At the other end of the spectrum is the model referred to as MG-MutSelHB, where each codon site has its own fixation factors, whereas all sites share the same mutational parameters (inspired by Halpern and Bruno 1998). Between these two extremes, the models denoted as MG-MutSelC20, MG-MutSelC40, and MG-MutSelC60 are empirical mixtures of 20, 40, and 60 components, which all share the same mutational parameters, but have distinct parameters governing fixation factors (as presented in Rodrigue et al. 2010). These mixture models provide a simple, crude, yet convenient compromise between the global homogeneous MG-MutSelYN model and the site-specific MG-MutSelHB model.

The importance of using reliable methods for computing Bayes factors in this mutation–selection modeling framework is immediately apparent from a simple simulation experiment. Using the alignment of 17 vertebrate globin genes taken from Yang et al. (2000) (and subsequently studied repeatedly by many authors), along with the tree topology used therein, we ran a plain MCMC sampling under the MG model. For 12 random draws from the posterior, we simulated new data sets conditional on those parameters (under the same MG model). For the resulting replicate data sets, we computed the natural log Bayes factors contrasting the MG-MutSelYN model to the MG model using both the HME approach and a TI model switch approach (see Rodrigue et al. 2008a). We repeated the calculations 10 times for each replicate, and the means and standard deviations of the natural log Bayes factors are shown in Table 1. Under these simulation conditions, based on the simple MG model, we expect that an appropriate model selection approach should favor the genuine MG model, thereby disfavoring the over-parameterized MG-MutSelYN model. Our results reveal that the Bayes factors computed with HME would lead one to believe that the MG and MG-MutSelYN are nearly equivalent (with natural log Bayes factors relatively close to 0). In contrast, with the TI calculations, the log Bayes factors are systematically negative, in strong favor of the simpler correct model. Although the standard deviations

Table 1. Mean and standard deviation (from 10 distinct calculations) of the natural logarithm of the Bayes factor in favor of the MG-MutSelYN model over the MG model on simulated data

	HME	TI
rep. 1	0.59 [2.85]	−16.04 [0.33]
rep. 2	0.04 [3.01]	−17.10 [0.29]
rep. 3	3.07 [3.11]	−12.96 [0.31]
rep. 4	−0.61 [2.72]	−17.22 [0.30]
rep. 5	0.13 [3.63]	−17.40 [0.24]
rep. 6	1.19 [2.68]	−17.70 [0.22]
rep. 7	2.41 [3.12]	−18.64 [0.30]
rep. 8	4.41 [4.02]	−19.53 [0.28]
rep. 9	−0.81 [3.03]	−18.62 [0.27]
rep. 10	1.15 [2.98]	−19.81 [0.28]
rep. 11	1.74 [2.70]	−17.51 [0.31]
rep. 12	−3.74 [2.97]	−14.86 [0.27]

Notes: A negative value corresponds to a Bayes factor in favor of the simpler MG model, which was used to simulate the data. Results are reported for the HME and TI.

obtained are greater for the HME than for the TI calculations, the problem with the HME does not appear to be one of a trivial Monte Carlo error.

Computational Speedup with Data Augmentation

The estimation of Bayes factors with TI in the mutation–selection codon substitution context is tractable under a model based on a single substitution matrix such as MG-MutSelYN (Rodrigue et al. 2008b), albeit still requiring weeks of computation for small single-gene data sets on typical desktop computers. The same TI-based estimation of Bayes factor under the site heterogeneous MG-MutSelC20, MG-MutSelC40, or MG-MutSelC60 models pushes the limits of tractability, requiring months of computing time on a modern desktop computer (Rodrigue et al. 2010). Although the computational demand could be reduced by using SS instead of TI, it would likely still not allow for the calculation of Bayes factors involving the MG-MutSelHB model, which specifies as many codon substitution matrices as there are codon sites in the alignments. Indeed, the MG-MutSelHB model remains essentially intractable within a full pruning-based MCMC sampling system, and the applicability of such rich mutation–selection models has only recently become feasible thanks to another methodological development receiving increasing attention in Bayesian phylogenetic contexts: data augmentation (e.g., Lartillot 2006; Mateiu and Rannala 2006; de Koning et al. 2010).

The basic idea of data augmentation is to generate (using a set of Gibbs sampling schemes, e.g., Nielsen 2002; Rodrigue et al. 2008b) a realization of the Markov substitution process conditional on a particular hypothesis vector θ, at the internal nodes and along the branches of the phylogeny, and conditional on the states in the alignment, for all sites (for a detailed introduction, see, e.g., Bollback 2005). We denote such a detailed substitution mapping as ξ. Then, Metropolis–Hastings (MH) updates (Metropolis et al. 1953; Hastings 1970) of the θ vector are performed based on an augmented likelihood function, written as $p(D, \xi|\theta, M)$. Following a round of such MH updates on θ, a new realization of the Markov substitution process is sampled as before, but this time conditional on the current θ, thereby initiating the next cycle. Altogether, the MCMC allows one to draw from the joint posterior distribution of θ and ξ:

$$p(\theta, \xi|D, M) = \frac{p(D, \xi|\theta, M)p(\theta|M)}{p(D|M)}. \quad (4)$$

The θ component of a sample drawn from Equation (4) is distributed as in Equation (3). Several variants of these approaches have been explored (e.g., Lartillot 2006; Mateiu and Rannala 2006; de Koning et al. 2010) and have demonstrated the potential to decrease computing time of plain MCMC samplers by several orders of magnitude under some of the more broadly used substitution models while enabling the study of substitution models that would otherwise remain intractable in full pruning-based sampling.

A difficulty for Bayesian model ranking with approaches employing the data augmentation scheme described above is that the best-performing TI method (the *model switch* approaches, Lartillot and Philippe 2006; Rodrigue et al. 2008a) and the best-performing SS method (the *generalized stepping-stone*, Fan et al. 2011) rely on sampling from distributions involving two distinctly defined substitution processes, via a "morphing" between the posteriors under the two models of interest. However, generating detailed substitution realizations along phylogenetic trees conditional on a morphing of posteriors implicating two substitution processes is not immediately obvious. In the terminology of Xie et al. (2011) and Fan et al. (2011), detailed data augmentation of the sort described above is not directly compatible with sampling from "power posteriors."

Using TI with Data Augmentation

There may be several different approaches to exploiting data augmentation within TI or SS calculations, depending on the specific type of sampling system. Previous works using data augmentation within a TI sampler have done so because no other alternatives were available (e.g., Rodrigue et al. 2006; Baele et al. 2008). We emphasize here that such a methodology is relevant even when the full pruning based sampler is technically possible. As an example, in Appendix 2, we describe one way of employing detailed data augmentation within a TI-MCMC framework for mutation–selection models. The basic idea is to perform a morphing of the substitution process itself. By using this method to analyze the same globin data set as above, we computed the natural log Bayes factors contrasting the mutation–selection models described in Appendix 1 against the MG model (used as a reference model). The approach yields precise results (Table 2), in far less computing time than the full pruning-based model switch TI. For instance, the full pruning-based TI performed in Rodrigue et al. (2010) to compute natural log Bayes factors of the MG-MutSelC20, MG-MutSelC40, and MG-MutSelC60 models against MG required nearly 4 months of computing time per run (at least two runs were performed to assess precision) and returned values of 236, 235, and 269 all within about one log unit. Here, on the other hand,

Table 2. Mean and standard deviation (from 10 distinct calculations) of the natural logarithm of the Bayes factor for the models considered, with MG used as a reference

	HME	DA-TI
MG-MutSelYN	68.70 [3.04]	44.31 [0.31]
MG-MutSelHB	349.20 [5.83]	159.25 [0.51]
MG-MutSelC20	249.00 [3.09]	236.27 [0.35]
MG-MutSelC40	272.10 [3.51]	256.81 [0.38]
MG-MutSelC60	284.50 [4.12]	269.04 [0.41]

Note: Results are reported for the HME and data augmentation–based thermodynamic integration (DA-TI).

with data augmentation and TI, the results for the same model comparisons (see Table 2) required less than 12 hours per run. This considerable speedup allowed us to perform 10 runs (5 bidirectional runs), which returned results all within half a log unit. Note that these 10 runs were performed to assess the precision of the calculations. However, in normal practice with the methods, performing only a few (say, one bidirectional pair of calculations) would likely be sufficient.

The Bayes factors computed with the HME and TI methods are once again quite different. Most strikingly, calculations done with the HME would lead one to elect MG-MutSelHB as the most appropriate model, whereas TI calculations give favor to the MG-MutSelC60 model. These results corroborate previous studies (Lartillot and Philippe 2006; Fan et al. 2011; Xie et al. 2011), indicating that Bayes factors computed using the HME tend to mislead one into selecting the most parameter-rich model. Note that this tendency is not a bias of the Bayes factor itself, but rather points to a serious issue in the way it is approximated with the HME.

Future Directions

Data augmentation–based posterior sampling is still relatively new to phylogenetic modeling contexts, and new methods are being proposed regularly (e.g., Pedersen and Jensen 2001; Nielsen 2002; Robinson et al. 2003; Rodrigue et al. 2008b; de Koning et al. 2010). Reliable methods for Bayesian model ranking, that is, excluding HME, have received less attention. Methods such as reversible jump sampling (Green 1995) or the Dickie-Savage ratio (Verdinelli and Wasserman 1995) have rarely been used in phylogenetics beyond the seminal papers introducing them, respectively, by Huelsenbeck et al. (2004) and Suchard et al. (2001). Each method poses its own set of technical challenges and limitations, and the basic thermodynamic method (Lartillot and Philippe 2006) is no exception, where the computational costs have likely discouraged many. Recent work by Xie et al. (2011) and Fan et al. (2011), however, could help revive this area.

The method we present in Appendix 2 also shows promise and is illustrative of a context-specific design of TI for the mutation–selection codon models. Our method, however, has a number of limitations. For instance, additional computer code needs to be written to accommodate the morphing between the particular substitution models of interest. Also, the data augmentation approach employed here has the weakness of working only on a fixed topology. This is mainly because no MH kernels for performing joint moves on topology and data augmentation have yet been proposed. One potential alternative would be to revert to full pruning-based moves on the tree topology, with calculations dispatched to multiple processors or to the processing units present on most graphic cards (Suchard and Rambaut 2009)—provided that the hardware improves to allow for double precision calculations.

Many possible sampling approaches can readily be envisaged, and future studies could be based on exploiting the advantages of various existing ideas, such as approximating Bayes factors via combinations of TI-based approaches (Lartillot and Philippe 2006; Fan et al. 2011; Xie et al. 2011), Savage-Dickie ratios (Suchard et al. 2001), and Laplace approximations (e.g., Rodrigue et al. 2007). As we have explored here, using data-augmentation ideas within such methodologies could be computationally fruitful, and approaches other than the full Gibbs updating employed here (e.g., de Koning et al. 2010) should also be investigated in such contexts. Altogether, through advances along these lines, there is hope to eventually achieve a more efficient and reliable computational framework for comparing hypotheses under a given model, as well as model ranking in Bayesian phylogenomic analysis.

Funding

This work was supported by the Natural Sciences and Engineering Research Council of Canada (N.R., S.A.-B.), the Canada Foundation for Innovation (S.A.-B.), and the Quebec Centre for Biodiversity Science (N.R.).

Acknowledgments

We wish to thank Nicolas Lartillot for discussions and comments on the manuscript, as well as Mark Holder, David Posada, Vladimir Minin, and an anonymous reviewer.

References

Alekseyenko A., Lee C., Suchard M. 2008. Wagner and Dollo: a stochastic duet by composing two parsimonious solos. Syst. Biol. 57:772–784.

Baele G., Peer Y.V., Vansteelandt S. 2008. A model-based approach to study nearest-neighbor influences reveals complex substitution patterns in non-coding sequences. Syst. Biol. 57:675–692.

Bleidorn C., Eeckhaut I., Podsiadlowski L., Schults N., McHugh D., Halanych K., Milinkovitch M., Tiedemann R. 2007. Mitochondrial genome and nuclear sequence data support Myzostomida as part of the annelid radiation. Mol. Biol. Evol. 24:1690–1701.

Bollback J.P. 2005. Posterior mapping and posterior predictive distributions. In: Nielsen R., editor. Statistical methods in molecular evolution. New York: Springer. p. 439–462.

Brown J., Lemmon A. 2007. The importance of data partitioning and the utility of Bayes factors in Bayesian phylogenetics. Syst. Biol. 56:643–655.

de Koning J.A.P., Gu W., Pollock D.D. 2010. Rapid likelihood analysis on large phylogenies using partial sampling of substitution histories. Mol. Biol. Evol. 27:249–265.

Drummond A.J., Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7:214.

Fan Y., Wu R., Chen M.-H., Kuo L., Lewis P.O. 2011. Choosing among partition models in Bayesian phylogenetics. Mol. Biol. Evol. 28: 523–532.

Green P.J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika. 82:711–732.

Halpern A.L., Bruno W.J. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol. Biol. Evol. 15:910–917.

Hampl V., Hug L., Leigh J.W., Dacks J.B., Lang B.F., Simpson A.G.B., Roger A. 2009. Phylogenomic analyses support the monophyly

of Excavata and resolve relationships among eukaryotic "supergroups". Proc. Natl. Acad. Sci. U. S. A. 106:3859–3864.
Hastings W.K. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika. 57:97–109.
Huelsenbeck J.P., Larget B., Alfaro M.E. 2004. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. Mol. Biol. Evol. 21:1123–1133.
Jeffreys H. 1935. Some tests of significance, treated by the theory of probability. Proc. Camb. Philos. Soc. 31:203–222.
Kass R., Raftery A. 1995. Bayes factors and model uncertainty. J. Am. Stat. Assoc. 90:773–795.
Lartillot N. 2006. Conjugate Gibbs sampling for Bayesian phylogenetic models. J. Comput. Biol. 13:1701–1722.
Lartillot N., Philippe H. 2006. Computing Bayes factors using thermodynamic integration. Syst. Biol. 55:195–207.
Le, S.Q., Gascuel O., Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics. 24:2317–2323.
Mateiu L., Rannala B. 2006. Inferring complex DNA substitution processes on phylogenies using uniformization and data augmentation. Syst. Biol. 55:259–269.
Metropolis S., Rosenbluth A.W., Rosenbluth M.N., Teller A.H., Teller E. 1953. Equation of state calculation by fast computing machines. J. Chem. Phys. 21:1087–1092.
Muse S.V., Gaut B.S. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitutions, with applications to chloroplast genome. Mol. Biol. Evol. 11:715–724.
Newton M.A., Raftery A.E. 1994. Approximating Bayesian inference with the weighted likelihood bootstrap. J. R. Stat. Soc. B. 56:3–48.
Nielsen R. 2002. Mapping mutations on phylogenies. Syst. Biol. 51: 729–739.
Pedersen A.-M.K., Jensen J.L. 2001. A dependent rates model and MCMC based methodology for the maximum likelihood analysis of sequences with overlapping reading frames. Mol. Biol. Evol. 18:763–776.
Praz C., Muller A., Danforth B., Griswold T., Widmer A., Dorn S. 2008. Phylogeny and biogeography of bees and the tribe Osmiini (Hymenoptera: Megachilidae). Mol. Phylogenet. Evol. 49:185–197.
Robinson D.M., Jones D.T., Kishino H., Goldman N., Thorne J.L. 2003. Protein evolution with dependence among codons due to tertiary structure. Mol. Biol. Evol. 20:1692–1704.
Rodrigue N., Kleinman C., Philippe H., Lartillot N. 2009. Computational methods for evaluating phylogenetic models of coding sequence evolution with dependence between codon. Mol. Biol. Evol. 26:1663–1676.
Rodrigue N., Philippe H., Lartillot N. 2006. Assessing site-interdependent phylogenetic models of sequence evolution. Mol. Biol. Evol. 23:1762–1775.
Rodrigue N., Philippe H., Lartillot N. 2007. Exploring fast computational strategies for probabilistic phylogenetic analysis. Syst. Biol. 56:711–726.
Rodrigue N., Philippe H., Lartillot N. 2008a. Bayesian comparisons of codon substitution models. Genetics. 180:1579–1591.
Rodrigue N., Philippe H., Lartillot N. 2008b. Uniformization for sampling realizations of Markov processes: applications to Bayesian implementations of codon substitution models. Bioinformatics. 24:56–62.
Rodrigue N., Philippe H., Lartillot N. 2010. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. Proc. Natl. Acad. Sci. U. S. A. 107:4629–4634.
Ronquist F., Huelsenbeck J.P. 2003. Mrbayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19:1572–1574.
Suchard M., Weiss R., Sinsheimer J. 2001. Bayesian selection of continuous-time Markov chain evolutionary models. Mol. Biol. Evol. 18:1001–1013.
Suchard M.A., Rambaut A. 2009. Many-core algorithms for statistical phylogenetics. Bioinformatics. 25:1370–1376.
Verdinelli I., Wasserman L. 1995. Computing Bayes factors using a generalization of the Savage-Dickey density ratio. J. Am. Stat. Assoc. 90:614–618.
Xie W., Lewis P.O., Fan Y., Kuo L., Chen M.-H. 2011. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. Syst. Biol. 60:150–160.
Yang Z., Nielsen R. 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol. Biol. Evol. 25:568–579.
Yang Z., Nielsen R., Goldman N., Pedersen A.-M.K. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics. 155:431–449.

APPENDIX 1: MUTATION–SELECTION MODELS

The basic substitution process on which we focus has infinitesimal rates from codon a to codon b given by:

$$Q_{ab} \propto \begin{cases} \rho_{a_c b_c} \varphi_{b_c} & \text{if } \mathcal{A}, \\ \rho_{a_c b_c} \varphi_{b_c} \frac{\beta S}{1-e^{-\beta S}} & \text{if } \mathcal{B}, \\ 0 & \text{otherwise,} \end{cases} \quad (5)$$

where:

$\mathcal{A}$: a and b are synonymous and differ only at codon position c;
$\mathcal{B}$: a and b are nonsynonymous and differ only at codon position c;

and where:

- $\rho = (\rho_{lm})_{1 \leqslant l,m \leqslant 4}$ is a set of symmetrical nucleotide relative exchangeability parameters, with the arbitrary constraint $\sum_{1 \leqslant l < m \leqslant 4} \rho_{lm} = 1$;
- $\varphi = (\varphi_l)_{1 \leqslant l \leqslant 4}$, with $\sum_{l=1}^{4} \varphi_l = 1$, represents a set of global nucleotide equilibrium propensities;
- β is a parameter to be used for the TI: when $\beta = 0$, the model amounts to a simple model without selection, which we refer to as MG (because it takes inspiration from Muse and Gaut 1994); when $\beta = 1$, we have the mutation–selection model of interest, denoted MG-MutSelYN (because it includes the mutation–selection approach used by Yang and Nielsen 2008);
- S is the scaled selection coefficient (scaled by twice the effective population size) associated with an amino acid replacing mutation. As described in Rodrigue et al. (2010), $S = \ln(\phi_{f(b)}/\phi_{f(a)})$, where $f(a)$ returns an index (ranging from 1 to 20) of the amino acid encoded by codon a, and $\phi = (\phi_n)_{1 \leqslant n \leqslant 20}$, with $\sum_{n=1}^{20} \phi_n = 1$, is a set amino acid preference parameters.

Note that when S or β (or their product) approaches 0, the fixation factor $\frac{\beta S}{1-e^{-\beta S}}$ approaches 1, but for numerical stability, we use the third-order Taylor approximation of the exponential term in the denominator in such cases. This leads to a fixation factor approximated as $\frac{1}{1-\beta S/2+\beta S^2/6}$.

We also use empirical mixtures models, designated as MG-MutSelC20, MG-MutSelC40, and MG-MutSelC60, constructed from 20, 40 and 60 different ϕ vectors, fixed to the values obtained by Le et al. (2008). The free parameters introduced by these models only consist of weights associated with the components of the mixture.

However, we use a sampling system based on an allocation variable, specifying a configuration in which each site is affiliated to one of the 20, 40, or 60 different components (e.g., see Le et al. 2008). Performing MH updates to the allocation variable implicitly integrates away the weight parameters.

Finally, we also explore a model in which each site has its own ϕ vector. Because it is inspired from Halpern and Bruno (1998), we refer to it as MG-MutSelHB.

We used the same priors described in Rodrigue et al. (2010). For the MG-MutSelHB model, not included in the latter study, we used independent flat *Dirichlet* priors on each vector (we did not explore a flexible hyperprior structure controlling the overall mean and variance of site-specific variables).

Appendix 2: Data Augmentation–based TI

We describe a TI method similar to that used in Rodrigue et al. (2006, 2009). In the following, we denote all parameters collectively by θ. The logarithm of the augmented likelihood function is a sum over all N codon sites in the alignment:

$$\ln p(D, \xi|\theta) = \sum_{i=1}^{N} \ln p(D_i, \xi_i|\theta), \tag{6}$$

where we have dropped the dependence on M (the overall form of the model) from the notation. The logarithm of the augmented likelihood for a particular codon site i is written as:

$$\ln p(D_i, \xi_i|\theta) = \ln p(s_i^{\text{root}}|\theta) + \sum_{j=1}^{2P-3} \ln p(s_{ij}, \xi_{ij}|s_{ij_{\text{up}}}, \theta), \tag{7}$$

where ξ_{ij} represents the substitution mapping at site i along the branch j (omitting index j signifies that the mapping over all branches), P is the number of sequences in the alignment, j_{up} is the parent node of node j, s_{ij} and $s_{ij_{\text{up}}}$ are the codon states at both ends of the branch for site i, and s_i^{root} is the codon state at site i of the sequence at the root node.

The derivative of the first term in Equation (7) with respect to β is given as:

$$\begin{aligned}\frac{\partial \ln p(s_i^{\text{root}}|\theta)}{\partial\beta} &= \frac{\partial}{\partial\beta} \ln \varphi_{s_{i_1}^{\text{root}}} \varphi_{s_{i_2}^{\text{root}}} \varphi_{s_{i_3}^{\text{root}}} e^{\beta \ln \phi_{f(s_i^{\text{root}})}} \\ &\quad - \frac{\partial}{\partial\beta} \ln \sum_{a=1}^{61} \varphi_{a_1} \varphi_{a_2} \varphi_{a_3} e^{\beta \ln \phi_{f(a)}} \qquad (8) \\ &= \ln \phi_{f(s_i^{\text{root}})} - \sum_{a=1}^{61} p(a|\theta) \ln \phi_{f(a)}. \qquad (9)\end{aligned}$$

Note that the second term in Equation (9) corresponds to an expectation taken with respect to the stationary distribution of the substitution process.

The derivative of the second term in Equation (7) is taken one branch at a time and is written as:

$$\begin{aligned}&\frac{\partial \ln p(s_{ij}, \xi_{ij}|s_{ij_{\text{up}}}, \theta)}{\partial\beta} \\ &= \left(\sum_{k=1}^{z_{ij}} \frac{\partial \ln Q_{s_{ijk-1}s_{ijk}}}{\partial\beta} - \frac{\partial (t_{ijk} - t_{ijk-1}) q_{s_{ijk-1}}}{\partial\beta} \right) \\ &\quad - \frac{\partial (\lambda_j - t_{ijz_j}) q_{s_{ijz_j}}}{\partial\beta},\end{aligned} \tag{10}$$

where:

- z_{ij} is the number of substitution events for site i between nodes j_{up} and j;
- t_{ijk} is the timing of the k^{th} event along branch j at site i, with $t_{ij0} = 0$;
- s_{ijk-1} and s_{ijk} are the codon states before and after event k along branch j for site i, with $s_{ij0} = s_{ij_{\text{up}}}$ and $s_{ijz_{ij}} = s_{ij}$;
- $q_a = \sum_{b \neq a} Q_{ab}$ is the *rate away* from codon state a;
- and λ_j is the length of branch j.

The derivative written in Equation (10) can be evaluated based on:

$$\frac{\partial \ln Q_{ab}}{\partial\beta} = \frac{e^{\beta S} - \beta S - 1}{\beta (e^{\beta S} - 1)} \tag{11}$$

and

$$\frac{\partial Q_{ab}}{\partial\beta} = \frac{\mu S - e^{-\beta S}(\mu S + \mu\beta S^2)}{e^{-2\beta S} - 2e^{-\beta S} + 1}, \tag{12}$$

where $\mu = \rho_{a_c b_c} \varphi_{b_c}$.

Again for reasons of numerical stability in cases where βS approaches 0 (in particular at one end of the TI where β is very small), we use third-order Taylor approximations.

The TI principle in this context consists in cycling over MH updates on all parameters of the model, except for β, which is instead updated deterministically using the quasi-static approach (Lartillot and Philippe 2006). This approach consists in starting the sampler with β=0, then adding a small increment $\delta\beta$ after a series of MH cycles over all other parameters (and auxiliary variables). The h^{th} draw from such a sample, $(\theta^{(h)}, \xi^{(h)})_{0 \leqslant h \leqslant K}$, is associated with β_h, where $\beta_0 = 0$, $\beta_K = 1$, and $\forall h \in [0, K]$, $\beta_{h+1} - \beta_h = \delta\beta$. The overall sample can be used to approximate the difference in marginal log likelihood at both ends of

the quasi-static procedure:

$$\ln \frac{p(D|\beta_K)}{p(D|\beta_0)} = \int_0^1 \left\langle \frac{\partial \ln p(D, \xi|\theta)}{\partial \beta} \right\rangle d\beta \quad (13)$$

$$\simeq \frac{1}{K}\left[\frac{1}{2}\frac{\partial \ln p(D, \xi^{(0)}|\theta^{(0)})}{\partial \beta} + \sum_{h=1}^{K-1} \frac{\partial \ln p(D, \xi^{(h)}|\theta^{(h)})}{\partial \beta} + \frac{1}{2}\frac{\partial \ln p(D, \xi^{(K)}|\theta^{(K)})}{\partial \beta}\right]. \quad (14)$$

Because the model at the beginning of the run, with $\beta = 0$, is equivalent to the basic MG model, and the model at the end of the run, with $\beta = 1$, is the target mutation–selection model, the approximation given in Equation (14) directly gives the thermodynamic log Bayes factor estimate between these two models (i.e., it constitutes a form of *model switch* TI).

We used sampling methods described elsewhere (e.g., Rodrigue et al. 2008b, 2009, 2010) and used $K = 10,000$ ($\delta\beta = 0.0001$). Moreover, we performed bidirectional integrations, which consist in running two independent calculations, one with β going from 0 to 1 and the other with β going from 1 to 0 (our 10-fold calculations using the proposed data augmentation–based TI actually consist of five bidirectional calculations).