Supporting Text

Parameter updates. We update a single parameter θ_i at a time using the independence sampler, a special case of the Metropolis–Hastings algorithm (1). A new value θ_i' is chosen from a proposal distribution $q(\theta_i'|\theta_{-i}, M, D)$, where M and D represent the current missing and observed data, respectively, and is accepted (replaces θ_i) with probability

$$\alpha = \min\left(1, \frac{q(\theta_i|\theta_{-i}, M, D)p(\theta_i'|\theta_{-i}, M, D)}{q(\theta_i'|\theta_{-i}, M, D)p(\theta_i|\theta_{-i}, M, D)}\right);$$

otherwise, θ_i remains unchanged. By Bayes' Theorem and our choice of prior, for any valid θ_i'

$$\frac{p(\theta_i'|\theta_{-i}, M, D)}{p(\theta_i|\theta_{-i}, M, D)} = \frac{p(M, D|\theta_i', \theta_{-i})p(\theta_i'|\theta_{-i})p(\theta_{-i})}{p(M, D|\theta_i, \theta_{-i})p(\theta_i|\theta_{-i})p(\theta_{-i})} = \frac{p(M, D|\theta_i', \theta_{-i})}{p(M, D|\theta_i, \theta_{-i})}.$$

For efficient sampling, $q(\theta_i'|\theta_{-i}, M, D)$ should be close to $p(\theta_i'|\theta_{-i}, M, D)$, with tails no lighter than those of p (2). With $D_i = \{X_{it} \,|\, t = m - k + 1, \ldots, m\}$, the components (M_1, D_1), (M_2, D_2), …, (M_n, D_n) form a Markov chain, and hence the Bernstein–von Mises Theorem extended to stochastic processes implies that the distribution $p(\theta_i'|\theta_{-i}, M, D)$ is close to a normal distribution centered at $\hat{\theta}_i$ (the maximum likelihood estimate of θ_i given θ_{-i}, M, and D) and having variance approximately equal to the inverse of the observed Fisher information $\mathcal{F}(\hat{\theta}_i) = -\frac{\partial^2}{\partial\theta_i^2}\log p(M, D, \theta_{-i}|\theta_i)\Big|_{\theta_i=\hat{\theta}_i}$ (3).

Under our discrete-time model of evolution, the log likelihood at $\theta = \{\lambda, \pi, \kappa\}$ is

$$\begin{aligned}\log p(M, D|\theta) \;=\; & \sum_{\substack{\sigma,b,\\ w,x,y\in\{\mathrm{A,C,G,T}\}}} \left[\begin{array}{r} N_{\sigma bwxyx}\log(1 - \sum_{z\neq x}\kappa_{b\sigma\tau_{wxy\to z}}\lambda_{\sigma\eta_b}(wxy\to z)) \\ + \sum_{z\neq x} N_{\sigma bwxyz}(\log\kappa_{b\sigma\tau_{wxy\to z}} + \log\lambda_{\sigma\eta_b}(wxy\to z))\end{array}\right] \\ & + \sum_{\substack{\rho,\\ v,w,x\in\{\mathrm{A,C,G,T}\}}} R_{\rho vwx}\log\pi_\rho(x|v, w),\end{aligned}$$

where $N_{\sigma bwxyz}$ is the number of occurrences in (M, D) of the base x mutating to z (if $z \neq x$) or not mutating (if $z = x$) in one time unit on branch b when x has neighboring bases w and y, and $R_{\rho vwx}$ is the number of root or newly inserted bases x preceded by v and w in sequence composition category ρ. Hence, the counts N and R are sufficient statistics, and we may analytically compute $\frac{p(M,D|\theta_i',\theta_{-i})}{p(M,D|\theta_i,\theta_{-i})}$ for θ_i any κ, λ, or π parameter, as well as $\hat{\pi}_\rho$ and $\mathcal{F}(\hat{\pi}_\rho)$ for each ρ, as a function of these counts. Computing $\hat{\theta}_i$ and $\mathcal{F}(\hat{\theta}_i)$ for the κ and λ parameters would require solving numerically, so

we instead analytically derive the analogous $\tilde{\theta}_i$ and $\mathcal{F}(\tilde{\theta}_i)$ for the continuous-time model as follows (note that $\hat{\pi}_\rho = \tilde{\pi}_\rho$ and $\mathcal{F}(\hat{\pi}_\rho) = \mathcal{F}(\tilde{\pi}_\rho)$).

Under the continuous-time model, the log likelihood (excluding terms for the π parameters) is

$$f(\theta|M,D) = \sum_{\substack{\sigma,b,\\ w,x,y\in\{\mathrm{A,C,G,T}\},\\ z\neq x}} \left[-\kappa_{b\sigma\tau_{wxy\to z}}\lambda_{\sigma\eta_b}(wxy\to z)T_{b\sigma wxy} + N_{\sigma bwxyz}(\log\kappa_{b\sigma\tau_{wxy\to z}} + \log\lambda_{\sigma\eta_b}(wxy\to z))\right],$$

where $T_{b\sigma wxy} = \sum_z N_{\sigma bwxyz}$ represents the total amount of time spent as base x in region type σ and branch b with neighbors w and y. Setting partial derivatives to 0 and solving, we have

$$\tilde{\kappa}_{b\sigma\tau} = \frac{\sum_{w,x,y,z\neq x\ :\ \tau_{wxy\to z}=\tau} N_{\sigma bwxyz}}{\sum_{w,x,y,z\neq x\ :\ \tau_{wxy\to z}=\tau} \lambda_{\sigma\eta_b}(wxy\to z)T_{b\sigma wxy}}$$

and

$$\tilde{\lambda}_{\sigma\eta}(wxy\to z) = \frac{\sum_{b\ :\ \eta_b=\eta} N_{\sigma bwxyz}}{\sum_{b\ :\ \eta_b=\eta} \kappa_{b\sigma\tau_{wxy\to z}}T_{b\sigma wxy}}.$$

From the second derivatives, we estimate variances

$$\mathcal{F}(\tilde{\kappa}_{b\sigma\tau})^{-1} = \frac{\sum_{w,x,y,z\neq x\ :\ \tau_{wxy\to z}=\tau} N_{\sigma bwxyz}}{\left(\sum_{w,x,y,z\neq x\ :\ \tau_{wxy\to z}=\tau} \lambda_{\sigma\eta_b}(wxy\to z)T_{b\sigma wxy}\right)^2}$$

and

$$\mathcal{F}(\tilde{\lambda}_{\sigma\eta}(wxy\to z))^{-1} = \frac{\sum_{b\ :\ \eta_b=\eta} N_{\sigma bwxyz}}{\left(\sum_{b\ :\ \eta_b=\eta} \kappa_{b\sigma\tau_{wxy\to z}}T_{b\sigma wxy}\right)^2}.$$

We now take for q a t distribution with four degrees of freedom centered at $\tilde{\theta}_i$ and scaled to have variance $\mathcal{F}(\tilde{\theta}_i)^{-1}$, which approximates the normal distribution but with heavier tails. This choice of proposal distribution yields acceptance rates for the analyses described of $\approx$92% for the branch scale parameters, $\approx$82% for the substitution rates, and $\approx$64% for the root transition probabilities.

Missing data updates. We update M_i using Gibbs sampling (1) (independence sampling with $q = p$, so that $\alpha = 1$) via the following algorithm. The target distribution is

$$p(M_i|\theta, M_{-i}, D) = \frac{p(M_i, M_{-i}, D|\theta)}{\sum_{M_i'} p(M_i', M_{-i}, D|\theta)} = \frac{\prod_{t=0}^{m} r_t(X_{i\beta_t}, X_{it})}{\sum_{y_0,\ldots,y_m} \prod_{t=0}^{m} r_t(y_{\beta_t}, y_t)},$$

where the summation $\sum_{y_0,\ldots,y_m}$ is over $y_t \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ if $X_{it} \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ but over $y_t \in \{\phi\}$ if $X_{it} = \phi$, and

$$\begin{aligned} r_t(y_{\beta_t}, y_t) &= \pi_{\rho_{it}}(y_t|X_{it}^{--}, X_{it}^{-})\pi_{\rho_{it}^{+}}(X_{it}^{+}|X_{it}^{-}, y_t)\pi_{\rho_{it}^{++}}(X_{it}^{++}|y_t, X_{it}^{+}) \\ &\quad \cdot \psi_{it}^{-}(X_{i\beta_t}^{--}X_{i\beta_t}^{-}y_{\beta_t} \to X_{it}^{-})\psi_{it}(X_{i\beta_t}^{-}y_{\beta_t}X_{i\beta_t}^{+} \to y_t)\psi_{it}^{+}(y_{\beta_t}X_{i\beta_t}^{+}X_{i\beta_t}^{++} \to X_{it}^{+}), \end{aligned}$$

where ρ_{it}^{+} and ρ_{it}^{++} are the sequence composition categories for X_{it}^{+} and X_{it}^{++}, respectively, and $\psi_{it}^{-} = \psi_{jt}$ and $\psi_{kt}^{-} = \psi_{jt}$, where j and k are the sequence position indices of X_{it}^{-} and X_{it}^{+}. The last equality follows from the fact that all factors not involving the ith position cancel.

For each internal tree position $t = 0, \ldots, m-k$, let $\mathcal{A}_t = \{s : \beta_s = t\}$. For $t = m-k+1, \ldots, m$, define $S_t(x) = 1$ if $x = X_{it}$ or $X_{it} = \mathrm{N}$ and $S_t(x) = 0$ otherwise. Iteratively, for $t = m-k, \ldots, 0$, define $S_t(x) = \prod_{s\in\mathcal{A}_t} \sum_y r_{is}(x, y)S_s(y)$. Then $S_t(x)/\sum_z S_t(z)$ is the probability of obtaining the portion of the D_i that are descendants of position t in the tree given $X_{it} = x$, θ, M_{-i}.

The value of X_{i0} is then sampled according to $p(X_{i0}|\theta, M_{-i}, D) = S_0(X_{i0})/\sum_z S_0(z)$. Given X_{it}, we sample X_{is} for each $s \in \mathcal{A}_t$ using $p(X_{is}|X_{it}, \theta, M_{-i}, D) = r_{is}(X_{it}, X_{is})S_s(X_{is})/\sum_z r_{is}(X_{it}, z)S_s(z)$. Iterating for $t = 0, \ldots, m-k$, we obtain a realization of M_i drawn from $p(M_i|\theta, M_{-i}, D)$.

Missing data components are updated in a randomly permuted order of all positions; when all positions have been updated, a new permutation is chosen.

Implementation details. Based on initial rate estimates obtained by using the DNAML program from the PHYLIP package (version 3.6b) with default parameters (4), each branch along the tree was divided into two or more discrete time units such that the average substitution rate per time unit is <0.005, resulting in a total of 357 time units within the tree.

For each analysis, initial branch lengths were estimated using PHYLIP (4); initial context-dependent substitution rates were estimated by parsimony; initial transition probabilities for the

root and inserted sequence distribution were estimated from the observed sequences; and an initial realization of the missing data was generated by randomly choosing a base at each sequence site of each tree position and then updating each missing data component three times without any parameter updates.

The Markov chain Monte Carlo (MCMC) was then run until each missing data component was updated 500 times. Between successive missing data component updates, with probability 1/500, all branch length and substitution rate parameters were updated, and with probability 1/20, all root distribution parameters were updated. After a burn-in period consisting of 40 additional updates of the missing data components, we sampled each time 1% of the missing data components were updated. There were, on average, at least 16 accepted updates of the parameters between samples.

Dataset. Low complexity regions were detected using DUST (R. L. Tatusov & D. J. Lipman, unpublished) with default parameters. CpG islands were detected (5) as maximally scoring segments of the sequence (using a scoring scheme that assigns CpG dinucleotides a score of 17 and all other dinucleotides -1) having scores 50 or greater, and for which at least 20% of the segment falls outside of all annotated repeats. Removed segments were replaced by three consecutive N's to reduce spurious neighbor effects between remaining adjacent positions. Sequence positions were marked as either transcribed or untranscribed according to the human annotations and as part of a repeat according to annotations for each species.

The results in the paper were obtained without any filtering of the sequence by alignment quality. To evaluate the effect misalignments may have on obscuring true substitution events or incorrectly suggesting their occurrence, the analysis was repeated using the following algorithm to reduce the number of misaligned bases, by iteratively accepting well-aligned regions in each sequence. First, the entire human sequence was marked as accepted. For each of the other sequences, all bases within a sliding window of size 14 were accepted if at least 11 bases within the window, including two consecutive bases at each end of the window, matched accepted bases in another sequence. This process was continually repeated for all windows and all sequences until no new windows were accepted. Then any parts of the sequences that were not accepted were masked out. Repeating the analyses with poorly aligned sequences removed in this manner did not qualitatively affect the conclusions, although branch lengths between distantly related species were decreased.

Defining substitution types. To determine the optimal partitioning into types for explaining variation among lineages, we carried out a weighted ANOVA, taking advantage of the fact that we have reliable variance and covariance estimates for the parameter estimates from the MCMC analysis. We assume a class of models in which differences among clades are explained by multiplicative shifts of substitution rates so that, for clade η, $\log \lambda_\eta(wxy{\rightarrow}z) = \log \lambda(wxy{\rightarrow}z) + d_{\eta\tau} + \varepsilon_{\eta,wxy\rightarrow z}$, where $\lambda_\eta(wxy{\rightarrow}z)$ is the clade-specific substitution rate estimated from the MCMC sample, $\lambda(wxy{\rightarrow}z)$ is the clade-independent component, $d_{\eta\tau}$ is the clade-specific shift for the substitution type τ containing $wxy{\rightarrow}z$, and $\sum_\eta d_{\eta\tau} = 0$. We assume departures from this model arise from uncertainty in the MCMC estimates, so that under weak regularity conditions the values $\{\varepsilon_{\eta,wxy\rightarrow z}\}_{\forall\,\eta,wxy\rightarrow z}$ asymptotically follow a multivariate normal distribution, with mean 0 and a covariance matrix estimated as $\mathrm{Cov}\{\log \lambda_\eta(wxy{\rightarrow}z)\}$ from the MCMC sample. We use this normal approximation to define log likelihoods for discriminating among possible values of $\lambda(wxy{\rightarrow}z)$ and $d_{\eta\tau}$.

Only untranscribed region rate matrices for the primate, rodent + rabbit, and carnivore + artiodactyl + horse clades were used in the ANOVA because of relatively large variances for the other branch group estimates. For the simplest model in this class, there is a single substitution type; this model yielded a log likelihood of $-3,232.4$. Defining six symmetric substitution types of the form NxN$\rightarrow z$ increased the log likelihood to $-1,161.0$. We then considered splittings of each of these six types into subtypes by using a recursive procedure that accepts a splitting if the Bonferroni corrected P value (computed assuming the χ^2 distribution for twice the log likelihood improvement and taking account of the full number of tests performed) does not exceed 0.001. This procedure yielded a division into the 14 types listed in Table 1, with a corresponding log likelihood of -319.6.

1. Tierney, L. (1994) *Ann. Stat.* **22,** 1701–1728.
2. Roberts, G. O. (1999) *J. Appl. Probab.* **36,** 1210–1217.
3. Heyde, C. C. & Johnstone, I. M. (1979) *J. R. Statist. Soc. Ser. B Methodol.* **41,** 184–189.
4. Felsenstein, J. (1989) *Cladistics* **5,** 164–166.
5. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001) *Nature* **409,** 860–921.