METHODOLOGY

bModelTest: Bayesian phylogenetic site model averaging and model comparison

Remco R Bouckaert[1,2,3*] and Alexei J Drummond[1,2]

*Correspondence:
remco@cs.auckland.ac.nz
alexei@cs.auckland.ac.nz
[1]Centre for Computational Evolution, University of Auckland, Auckland, NZ
Full list of author information is available at the end of the article

Abstract

Background: Reconstructing phylogenies through Bayesian methods has many benefits, which include providing a mathematically sound framework, providing realistic estimates of uncertainty and being able to incorporate different sources of information based on formal principles. Bayesian phylogenetic analyses are popular for interpreting nucleotide sequence data, however for such studies one needs to specify a site model and associated substitution model. Often, the parameters of the site model is of no interest and an ad-hoc or additional likelihood based analysis is used to select a single site model.

Results: bModelTest allows for a Bayesian approach to inferring and marginalizing site models in a phylogenetic analysis. It is based on trans-dimensional Markov chain Monte Carlo (MCMC) proposals that allow switching between substitution models as well as estimating the posterior support for gamma-distributed rate heterogeneity, a proportion of invariable sites and unequal base frequencies. The model can be used with the full set of time-reversible models on nucleotides, but we also introduce and demonstrate the use of two subsets of time-reversible substitution models.

Conclusion: With the new method the site model can be inferred (and marginalized) during the MCMC analysis and does not need to be pre-determined, as is now often the case in practice, by likelihood-based methods. The method is implemented in the bModelTest package of the popular BEAST 2 software, which is open source, licensed under the GNU Lesser General Public License and allows joint site model and tree inference under a wide range of models.

Keywords: Model averaging; Model selection; Model comparison; Statistical phylogenetics; ModelTest; Phylogenetic model averaging; Phylogenetic model comparison; Substitution model; Site model

Background

One of the choices that needs to be made when performing a Bayesian phylogenetic analysis is which site model to use. A common approach is to use a likelihood-based method like ModelTest [1], jModelTest [2], or jModelTest2 [3] to determine the *site model*. The site model is comprised of (i) a substitution model defining the relative rates of different classes of substitutions and (ii) a model of rate heterogeneity across sites which may include a gamma distribution [4] and/or a proportion of invariable sites [5, 6]. The site model recommended by such likelihood-based method is then often used in a subsequent Bayesian phylogenetic analysis. This analysis framework introduces a certain circularity, as the original model selection step requires a phylogeny, which is usually estimated by a simplistic approach. Also, by forcing the subsequent Bayesian phylogenetic analysis to condition on the selected

site model, the uncertainty in the site model can't be incorporated into the uncertainty in the phylogenetic posterior distribution. A more statistically rigorous and elegant method is to co-estimate the site model and the phylogeny in a single Bayesian analysis, thus alleviating these issues.

One way to select substitution models for nucleotide sequences is to use reversible jump between all possible reversible models [7], or just a nested set of models [8]. An alternative is to use stochastic Bayesian variable selection [9], though this does not address whether to use gamma rate heterogeneity or invariable sites. Wu et al. [10] use reversible jump for substitution models and furthermore select for each site whether to use gamma rate heterogeneity or not. Since the method divides sites among a set of substitution models, it does not address invariable sites, and only considers a very limited set of five (K80, F81, HKY85, TN93, and GTR) substitution models.

We introduce a method which combines model averaging over substitution models with model averaging of the parameters governing rate heterogeneity across sites using reversible jump. Whether one considers the method to be selecting the site model, or averaging over (marginalizing over) site models depends on which random variables are viewed as parameters of interest and which are viewed as nuisance parameters. If the phylogeny is viewed as the parameter of interest, then bModelTest provides estimates of the phylogeny averaged over site models. Alternatively if the site model is of interest, then bModelTest can be used to select the site model averaged over phylogenies. These are matters of post-processing of the MCMC output, and it is also possible to consider the interaction of phylogeny and site models. For example one could construct phylogeny estimates conditional on different features of the site model from the results of a single MCMC analysis.

The method is implemented in the bModelTest package of BEAST 2 [11] with GUI support for BEAUti making it easy to use. It is open source and available under LGPL licence. Source code, installation instructions and documentation can be found at `https://github.com/BEAST2-Dev/bModelTest`.

Methods

All time-reversible nucleotide models can be represented by a 4×4 instantaneous rate matrix:

$$Q = \begin{pmatrix} - & \pi_C r_{ac} & \pi_G r_{ag} & \pi_T r_{at} \\ \pi_A r_{ac} & - & \pi_G r_{cg} & \pi_T r_{ct} \\ \pi_A r_{ag} & \pi_C r_{cg} & - & \pi_T r_{gt} \\ \pi_A r_{at} & \pi_C r_{ct} & \pi_G r_{gt} & - \end{pmatrix},$$

with six rate parameters r_{ac}, r_{ag}, r_{at}, r_{cg}, r_{ct} and r_{gt} and four parameters describing the equilibrium base frequencies $\Pi = (\pi_A, \pi_C, \pi_G, \pi_T)$. A particular restriction on the rate parameters can conveniently be represented by a six figure model number where each of the six numbers corresponds to one of the six rates in the alphabetic order listed above. Rates that are constrained to be the same, have the same integer at their positions in the model number. For example, model 123456 corresponds to a model where all rates are independent, named the general time reversible (GTR) model [12]. Model 121121 corresponds to the HKY model [13] in which rates form

two groups labelled transversions ($1 : r_{ac} = r_{at} = r_{cg} = r_{gt}$) and transitions ($2 : r_{ag} = r_{ct}$). By convention, the lowest possible number representing a model is used, so even though 646646 and 212212 represent HKY, we only use 121121.

There are 203 reversible models in total [7]. However, it is well known that transitions (A↔C, and G↔T substitutions) are more likely than tranversions (the other substitutions) [14, 15]. Hence grouping transition rates with transversion rates is often not appropriate and these rates should be treated differently. We can restrict the set of substitution models that allow grouping only within transitions and within transversions, with the exception of model 111111, where all rates are grouped. This reduces the 203 models to 31 models (see Figure 1 and details in Appendix). Alternatively, if one is interested in using named models, we can restrict further to include only Jukes Cantor [16, 17] (111111), HKY [13] (121121), TN93 [18] (121131), K81 [19] (123321), TIM [20] (123341), TVM [20] (123421),and GTR [12] (123456). However, to facilitate stepping between TIM and GTR during the MCMC (see proposals below) we like to use nested models, and models 123345 and 123451 provide intermediates between TIM and GTR, leaving us with a set of 9 models (Figure 1).

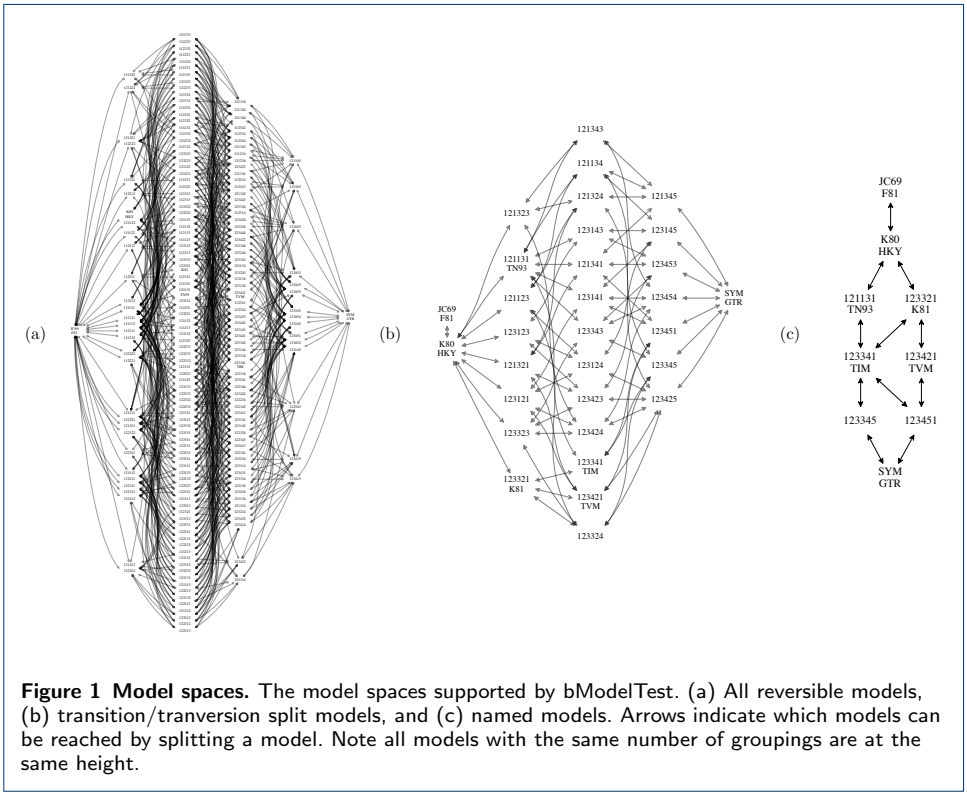

Figure 1 Model spaces. The model spaces supported by bModelTest. (a) All reversible models, (b) transition/tranversion split models, and (c) named models. Arrows indicate which models can be reached by splitting a model. Note all models with the same number of groupings are at the same height.

The state space consists of the following parameters:

- the model number M,
- a variable size rate parameter (depending on model number) R,
- a binary variable to indicate whether 1 or $k > 1$ non-zero rate categories should be used,
- a shape parameter α, used for gamma rate heterogeneity when there are $k > 1$ rate categories,

- a binary variable to indicate whether or not a category for invariable sites should be used,
- the proportion of invariable sites p_{inv},

Rates r_{ac}, r_{ag}, r_{at}, r_{cg}, r_{ct} and r_{gt} are determined from the model number M and rate parameter R. Further, we restrict R such that the sum of the six rates $\sum r_{..}$ equals 6 in order to ensure identifiability. This is implemented by starting each rate with value 1, and ensuring proposals keep the sum of rates in (see details on proposals below).

Prior

By default, bModelTest uses the flat Dirichlet prior on rates from [7]. From empirical studies [14, 15], we know that transition rates tend to be higher than transversion rates. It makes sense to encode this information in our prior and bModelTest allows for rates to get a different prior on transition rates (default log normal with mean 1 and standard deviation of 1.25 for the log rates) and transversion rates (default exponential with mean 1 for the rates).

An obvious choice for the prior on models is to use a uniform prior over all valid models. As Figure 1 shows, there are many more models with 3 parameters than with 1. An alternative allowed in bModelTest is to use a uniform prior on the number of parameters in the model. In that case, Jukes Cantor and GTR get a prior probability of 1/6, since these are the only models with 0 and 5 degrees of freedom respectively. Depending on the model set, a much lower probability is assigned to each of the individual models such that the total prior probability summed over models with k parameters, $p(k) = 1/6$ for $k \in \{0, 1, 2, 3, 4, 5\}$.

For frequencies a Dirichlet(4,4,4,4) prior is used, reflecting our believe that frequencies over nucleotides tend to be fairly evenly distributed, but allowing a 2.2% chance for a frequency to be under 0.05. For p_{inv} a Beta(4,1) prior on the interval $(0, 1)$ is used giving a mean of 0.2 and for α an exponential with a mean 1. These priors only affect the posterior when the respective binary indicator is 1.

MCMC proposals

The probability of acceptance of a (possibly trans-dimensional) proposal [21] is

$$\min\{1, \text{posterior ratio} \times \text{proposal ratio} \times \text{Jacobian}\}$$

where the posterior ratio is the posterior of the proposed state S' divided by that of the current state S, the proposal ratio the probability of moving from S to S' divided by the probability of moving back from S' to S, and the Jacobian is the determinant of the matrix of partial derivatives of the parameters in the proposed state with respect to that of the current state [21].

Model merge/split proposal

For splitting (or merging) substitution models, suppose we start with a model M. To determine the proposed model M', we randomly select one of the child (or parent) nodes in the graph (as shown in Figure 1). This is in contrast to the approach of Huelsenbeck *et al* [7], in which first a group is randomly selected, then a subgrouping

is randomly created. Our graph-based method is easier to generalise to other model sets (e.g. the one used in [22]). If there are no candidates to split (that is, model $M = 123456$ is GTR) the proposal returns the current state (this proposal is important to guarantee uniform sampling of models). Likewise, when attempting to merge model $M = 111111$, the current state is proposed ($M' = 111111$). Let r be the rate of the group to be split. We have to generate two rates r_i and r_j for the split into groups of size n_i and n_j. To ensure rates sum to 6, we select u uniformly from the interval $(-n_i r, n_j r)$ and set $r_i = r + u/n_i$ and $r_j = r - u/n_j$.

For a merge proposal, the rate of the merged group r from two split groups i and j with sizes n_i and n_j, as well as rates r_i and r_j is calculated as $r = \frac{n_i r_i + n_j r_j}{n_i + n_j}$.

When we select merge and split moves with equal probability, the proposal ratio for splitting becomes

$$\frac{\frac{1}{|M'_{merge}|}}{\frac{1}{|M_{split}|}} \frac{1}{r(n_i + n_j)}$$

where $|M_{split}|$ (and $|M'_{merge}|$) is the number of possible candidates to split (and merge) into from model M (and M' respectively). The proposal ratio for merging is

$$\frac{\frac{1}{|M'_{split}|}}{\frac{1}{|M_{merge}|}} r(n_i + n_j).$$

The Jacobian for splitting is $\frac{n_i + n_j}{n_i n_j}$ and for merging it is $\frac{n_i n_j}{n_i + n_j}$.

Rate exchange proposal

The rate exchange proposal randomly selects two groups, and exchanges a random amount such that the condition that all six rates sum to 6 is met. A random number is selected from the interval $[0, \delta]$ where δ is a tuning parameter of the proposal (δ is automatically optimized to achieve the desired acceptance probability for the data during the MCMC chain). Let n_i, r_i, n_j and r_j as before, then the new rates are $r'_i = r_i - u$ and $r'_j = r_j + u\frac{n_i}{n_j}$. The proposal fails when $r'_i < 0$.

The proposal ratio as well as the Jacobian are 1.

Birth/death proposal

Birth and death proposals set or unset the category count flag and sample a new value for α from the prior when the flag is set. The proposal ratio is $d(\alpha')$ for birth and $1/d(\alpha)$ for death where $d(.)$ is the density used to sample from (by default an exponential density with a mean of 1).

Likewise for setting the indicator flag to include a proportion of invariable sites and sampling p_{inv} from the prior. The Jacobian is 1 for all these proposals.

Scale proposal

For the α, we use the standard scale operator in BEAST 2 [11], adapted so it only samples if the category count flag is set for α. Likewise, for $pInv$ this scale operator is used, but only if the indicator flag to include a proportion of invariable sites is set.

Results and Discussion

Since implementation of the split/merge and rate exchange proposals is not straightforward, nor is derivation of the proposal ratio and Jacobian, unit tests were written to guarantee their correctness and lack of bias in proposals (available on https://github.com/BEAST2-Dev/bModelTest).

To validate the method we performed a simulation study by drawing site models from the prior, then used these models to generate sequence data of 10K sites length on a tree (in Newick (A:0.2,(B:0.15,C:0.15):0.05)) with three taxa under a strict clock. The data was analysed using a Yule tree prior, a strict clock and bModelTest as site model with uniform prior over models and exponential with mean one for tranversions and log-normal with mean one and variance 1.25 for transition rates. A hundred alignments were generated with gamma rate heterogeneity and a hundred without rate heterogeneity using a BEASTShell [23] script. Invariant sites can be generated in the process and are left in the alignment.

Comparing the model used to generate the alignments with inferred models is best done by comparing the individual rates of these models. Figure 2 shows the rate estimates for the six rates agains the rates used to generate the data. Clearly, there is a high correlation between the estimated rates and the ones used to generate ($R^2 > 0.99$ for all rates). Results were similar with and without rate heterogeneity. Note values for rates AG and CT (middle panels) tend to be higher than the transversion rates due to the prior they are drawn from.

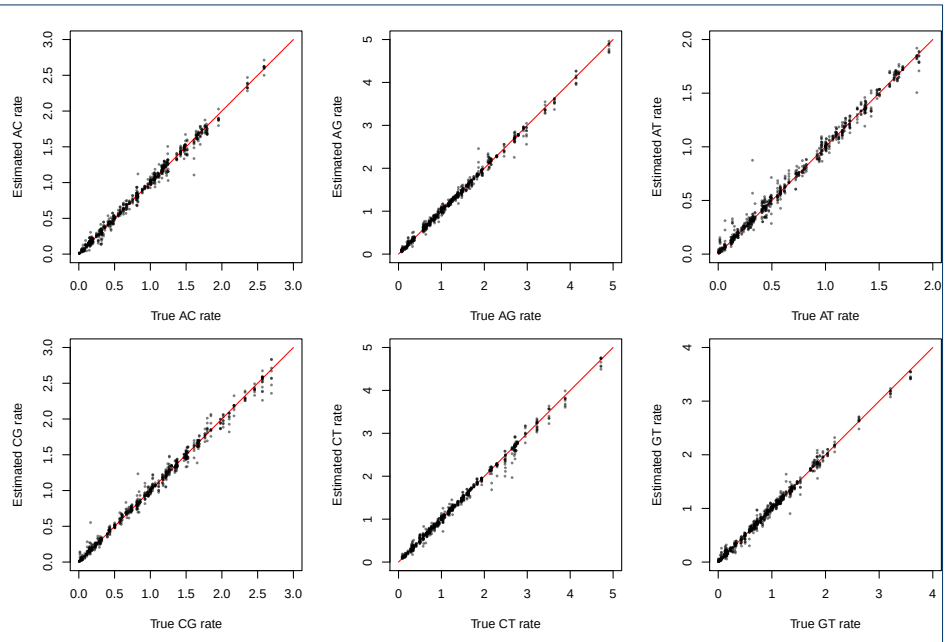

Figure 2 Accuracy of estimated substitution rates. True rates (horizontal) against estimated rates (vertical) in simulated data. In reading order, rate AC, AG, AT, CG, CT and GT.

Table 1 summarises coverage of the various parameters in the model, which is defined as the number of experiments where the 95% HPD of the parameter estimate contains the value of the parameter used to generate the data. On average, one would expect the coverage to be 95% if simulations are drawn from the prior [24].

The rows in the table show the four different models of rate heterogeneity among sites; plain means a single category without gamma or invariable sites, +G for discrete gamma rate categories, +I for two categories, one being invariable, and +G+I for discrete gamma rate categories and one invariable category. Furthermore, the experiment was run estimating whether base frequencies were equal or not. The first four rows are for data simulated with equal frequencies, the latter four with unequal frequencies. The last row shows results averaged over all 800 experiments.

		rate coverage						mean	Subst. Model
Freqs	Site Model	AC	AG	AT	CG	CT	GT	rate	coverage
equal	plain	93	97	94	96	95	95	95	98
equal	+G	91	95	93	93	95	93	93.3	97
equal	+I	92	94	94	95	93	94	93.6	96
equal	+G+I	89	96	95	94	95	95	94	98
unequal	plain	96	95	96	97	93	96	95.5	96
unequal	+G	95	94	94	94	96	96	94.8	98
unequal	+I	89	94	95	95	93	95	93.5	93
unequal	+G+I	97	94	94	93	93	96	94.5	97
Mean		94.25	94.25	94.75	94.75	93.75	95.75	94.6	96

					frequency	frequency coverage			
Freqs	Site Model	Site Model coverage	α	p_{inv}	coverage	A	C	G	T
equal	plain	100			100	91	95	99	95
equal	+G	96	94		100	92	94	98	97
equal	+I	98		95	100	91	93	98	96
equal	+G+I	99	89	88	100	91	93	98	95
unequal	plain	100			100	92	95	97	96
unequal	+G	97	94		100	97	92	92	98
unequal	+I	98		92	100	95	94	94	89
unequal	+G+I	100	93	91	100	99	96	96	98
Mean		98.75	93.50	91.50	100.00	95.75	94.25	94.75	95.25

Table 1 Coverage summary for simulation study. The first column lists the frequency and site models used to generate the data, and the last row is the mean coverage over all 800 runs. Coverage for rate parameters and frequencies is defined as the number of replicate simulations in which the true parameter value was contained in the estimated 95% HPD interval. The mean rate column contains the coverage averaged over all six rate coverage columns (i.e. the proportion of the 600 parameter estimates whose values were contained in their respective 95% HPD intervals. For details of substitution model coverage see text. The site model coverage is the number of replicate simulations that contained the correct model specification for rate heterogeneity across sites in the 95% credible set of models. Columns α and p_{inv} are coverages of the shape and proportion invariable parameter *conditioned on sampling from the true site model.*

Coverage of rate estimates and frequencies are as expected, as shown in the table. Substitution model coverage is measured by first creating the 95% credible set of models for each simulation and then counting how often the model used to generate the data was part of the 95% credible set. The 95% credible set is the smallest set of models having total posterior probability ≥ 0.95. As Table 1 shows, model coverage is as expected (Subst. Model coverage column). The situation with gamma shape parameter estimates and proportion of invariable sites is not as straightforward as for the relative rates of the substitution process. The site model coverage can be measured in a similar fashion: the site model coverage column shows how often the 95% credible sets for the four different site models (plain, +G. +I and +G+I) contains the true model used to generate the data. The coverage is as expected. When looking at how well the shape parameter (α column in Table 1) and the proportion invariable sites (p_{inv} column in the table) is estimated, we calculated the 95% HPD intervals for that part of the trace where the true site model was sampled. Coverage is as expected when only gamma rate heterogeneity is used, or when only a proportion of invariable sites is used, but when both are used an interaction between the two site model categories appears to slightly reduce the coverage of both parameters. In these experiments the coverage for the frequency

estimates for the individual nucleotides was as expected. In summary, the statistical performance of the model is as expected for almost all parameters except for the case where gamma and a proportion of invariable sites are used due to their interaction as discussed further below.

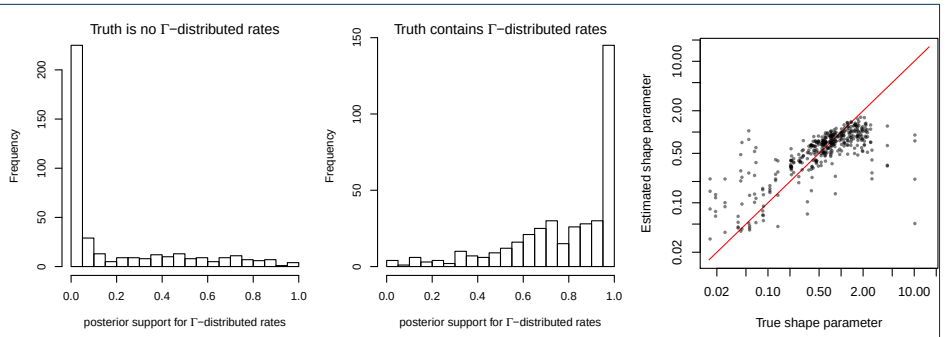

Figure 3 Accuracy of inference of rate heterogeneity across sites. Posterior probability for inclusion of gamma rate heterogeneity when the data is generated without (left) and with (middle) rate heterogeneity. Right, True gamma shape parameter (horizontal) against estimated shape parameter (vertical) when rate heterogeneity is used to generate the data.

Figure 3 shows histograms with the proportion of time gamma shapes were used during the MCMC run. When data was generated without gamma shapes, gamma rate heterogeneity was not used most of the time (left of Figure 3), while gamma rates were used for most of the analyses most of the time when gamma rate heterogeneity was present (middle of Figure 3).[1] When rate heterogeneity was present, shape estimates were fairly close to the ones used to generate the data (right of Figure 3). However, there were quite a few outliers, especially when the shape parameter was high (although this is harder to see on a log-log plot which was used here because of the uneven distribution of true values). This can happen due to the fact that when the gamma shape is small, a large proportion of sites gets a very low rate, and may be invariant, so that the invariable category can model those instances. The mean number of invariant sites was 6083 when no rate heterogeneity was used, while it was 6907 when rate heterogeneity was used, a difference of about 8% of the sites.

Figure 4 shows similar plots as Figure 3 but for the proportion of invariable sites. [2] Empirically for the parameters that we used for our simulations, it appears that if there are less than 60% invariant sites, adding a category to model them does not give a much better fit.

When a proportion of invariable sites was included in the simulation, there was a high correlation between the true proportion and the estimated proportion of invariable sites.

[1] Estimated shape parameters only take values of the shape parameter in account in the posterior sample where gamma rate heterogeneity was present.
[2] The estimated proportion of invariable sites only take values of the parameter in account in the posterior sample where the invariant category was present.

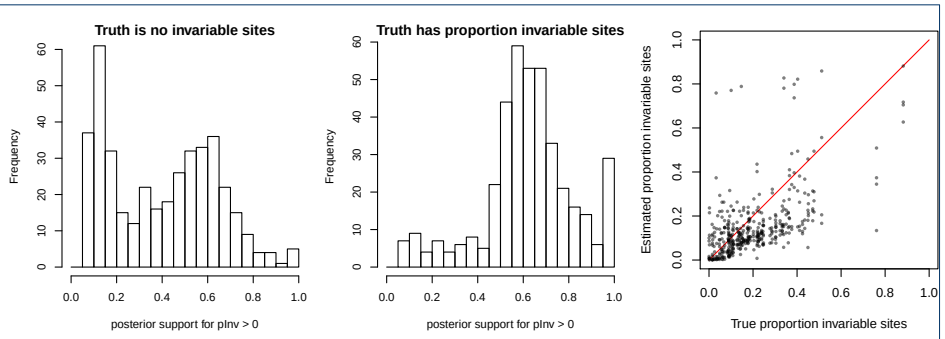

Figure 4 Accuracy of inference of proportion of invariant sites. Posterior probability for inclusion of a proportion of invariant sites when the data is generated without (left) and with (middle) invariant sites. Right, empirical proportion invariant in alignment (horizontal) against estimated proportion of invariant sites (vertical) when rate heterogeneity is used to generate the data.

Comparison with jModelTest on real data

To compare the application of bModelTest to jModelTest version 2.1.10 [3] (with settings `-f -i -g 4 -s 11 -AIC -a`) we applied both to two real datasets. The first data set used was an alignment from 12 primate species [25] (available from BEAST 2 as file examples/nexus/Primates.nex) containing 898 sites. In this case the model recommended by jModelTest was TPM2uf+G and the substitution model TPM2 (=121323) has the highest posterior probability using bModelTest (21.12% see Appendix for full list of supported models) when empirical frequencies are used. However, when frequencies are allowed to be estimated, HKY has highest support (16.19%), while TPM2 still has good support (10.25%) after model 121123 (14.09% support). So, using a maximum likelihood approach (jModelTest and/or empirical frequencies) makes a substantial difference in the substitution model being preferred. Figure 5 left shows the support for all models, and it shows that the 95% credible set is quite large for the primate data. Figure 5 middle and 5 right show correlation between substitution model rates. The former shows correlation between transversion rate AC (horizontally) and transition rate AG (vertically). One would not expect much correlation between these rates since the model coverage image shows there is little support for these rates to be shared. However, since HKY is supported to a large extent and the rates are constrained to sum to 6, any proposed change in a transition rate requires an opposite change in transversion rates in order for the sum to remain 6. So, when sampling HKY, there is a linear relation between transition and transversion rates, which faintly shows up in the Figure 5 (middle). Figure 5 (right) shows the correlation between transversion rates AC and AT. Since they are close to each other, a large proportion of the time rate AC and AT are linked, which shows up as a dense set of points on the AC=AT line.

The second data set used was an alignment of 31 sequences of 9030 sites of coding hepatitis C virus (HCV) from [26]. It was split into two partitions, the first containing codon 1 and 2 positions (6020 sites) and the second all codon 3 positions (3010 sites). Figure 6 left show the model distributions for the first partition at the top and second at the bottom. The 95% credible sets contain just 7 and 6 models respectively, much smaller than those for the primate data as one would expect from using longer, more informative sequences. Note that the models pre-

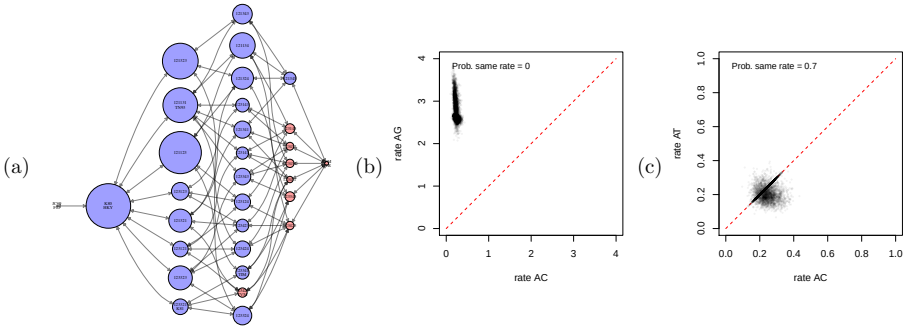

Figure 5 Posterior inference on primate data. Model distribution for primate data using the transition/tranversion split models (left). Numbers on x-axis correspond to models in Appendix. The middle panel plots rates $A \leftrightarrow C$ versus $A \leftrightarrow G$ (middle) and the right panel plots $A \leftrightarrow C$ versus $A \leftrightarrow T$.

ferred for the first partition have transition parameters split while for the second partition models where partitions are shared have more support. resulting in quite distinct model coverage images. For the first partition, bModelTest recommends TIM2+I+G. TIM2 is model 121343, the model with highest support according to bModelTest, as shown in Figure 6. For the second partition, jModelTest recommends GTR+G, and though GTR is in the 95% credible set, it gets much less support than TVM, even though TVM was considered by jModelTest. Again, we see a substantial difference in likelihood and Bayesian approaches. The correlation between transition rates $A \leftrightarrow G$ and $C \leftrightarrow T$ as well as between two tranversion rates $A \leftrightarrow C$ and $A \leftrightarrow T$ are shown in Figures 6 top middle and right for the first partition and Figures 6 bottom middle and right for the second. The transition rates $A \leftrightarrow G$ and $C \leftrightarrow T$ have a posterior probability of being the same of 0.024 in the first partition, whereas the posterior probability is 0.66 in the second partition containing only 3rd positions of the codons. This leads to most models for the first partition distinguishing between $A \leftrightarrow G$ and $C \leftrightarrow T$, while for the second partition most models share these rates. For the two transversion rates $A \leftrightarrow C$ and $A \leftrightarrow T$ the partitions display the opposite relationship, with the second partition preferring to distinguish them. As a result, overall the two partitions only have one model in common in their respective 95% credible sets, but that model (GTR) has quite low support from both partitions.

Implementation details

The calculation of the tree likelihood typically consumes the bulk ($\gg 90\%$) of computational time. Note that for a category with invariable sites, the rate is zero, hence only sites that are invariant (allowing for missing data) contribute to the tree likelihood. The contribution is 1 for those sites for any tree and for any parameter setting, so by counting the number of invariant sites, the tree likelihood can be calculated in constant time. Switching between with and without gamma rate heterogeneity means switching between one and k rate categories, which requires k time as much calculation. Having two tree likelihood objects, one for each of these two scenarios, and a switch object that selects the one required allows use of the

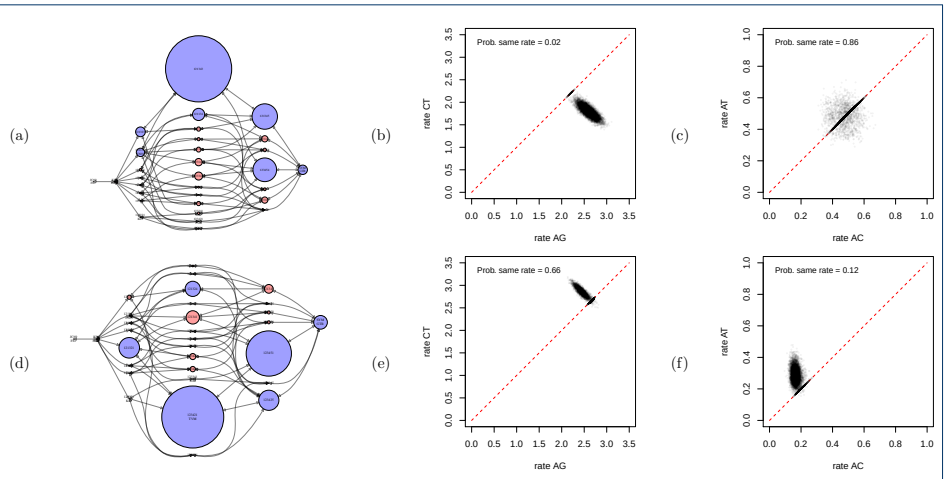

Figure 6 Posterior inference on HCV data. Like Figure 5, but the data is split into two partitions, the first containing codon positions 1+2 (panel a, b and c) and second containing codon position 3 (panel d, e and f). The partitions support distinctly different site models. The left panels show the posterior distribution over models, the middle panel plots transition rates $A \leftrightarrow G$ versus $C \leftrightarrow T$, and the right panel plots tranversion rates $A \leftrightarrow C$ versus $A \leftrightarrow T$.

BEAST 2 updating mechanism [9] so that only the tree likelihood that needs updating is performing calculations. So, jModelTest and bModelTest can, but do not necessarily agree on the most appropriate model to use.

Conclusions

bModelTest is a BEAST 2 package which can be used in any analysis where trees are estimated based on nucleotide sequences, such as multi-species coalescent analysis [27, 28], various forms of phylogeographical analyses, sampled ancestor analysis [29], demographic reconstruction using coalescent [30], birth death skyline analysis [31], *et cetera*. The GUI support provided through BEAUti makes it easy to set up an analysis with the bModelTest site model: just select bModelTest instead of the default gamma site model from the combo box in the site model panel.

bModelTest allows estimation of the site model using a full Bayesian approach, without the need to rely on non-Bayesian tools for selecting the site model.

List of abbreviations

BEAST: Bayesian evolutionary analysis by sampling trees
GTR: general time reversible
GUI: graphical user interface
HCV: hepatitis C virus
HPD: highest probability density
LGPL: Lesser General Public License
MCMC: Markov chain Monte Carlo

Funding

This work was funded by a Rutherford fellowship (http://www.royalsociety.org.nz/programmes/funds/rutherford-discovery/) from the Royal Society of New Zealand awarded to Prof. Alexei Drummond.

Author details

[1]Centre for Computational Evolution, University of Auckland, Auckland, NZ. [2]Department of Computer Science, University of Auckland, Auckland, NZ. [3]Max Planck Institute for the Science of Human History, Jena, Germany.

References

1. Posada, D., Crandall, K.A.: Modeltest: testing the model of dna substitution. Bioinformatics **14**(9), 817–818 (1998)
2. Posada, D.: jModelTest: phylogenetic model averaging. Molecular biology and evolution **25**(7), 1253–1256 (2008)
3. Darriba, D., Taboada, G.L., Doallo, R., Posada, D.: jModelTest 2: more models, new heuristics and parallel computing. Nature methods **9**(8), 772–772 (2012)
4. Yang, Z.: Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. Journal of Molecular Evolution **39**(3), 306–314 (1994). doi:10.1007/BF00160154
5. Gu, X., Fu, Y.X., Li, W.H.: Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. Mol Biol Evol **12**(4), 546–57 (1995)
6. Waddell, P., Penny, D.: Evolutionary trees of apes and humans from DNA sequences. In: Lock, A.J., Peters, C.R. (eds.) Handbook of Symbolic Evolution, pp. 53–73. Clarendon Press, Oxford., ??? (1996)
7. Huelsenbeck, J.P., Larget, B., Alfaro, M.E.: Bayesian phylogenetic model selection using reversible jump markov chain monte carlo. Molecular Biology and Evolution **21**(6), 1123–1133 (2004). doi:10.1093/molbev/msh123
8. Bouckaert, R.R., Alvarado-Mora, M., Rebello Pinho, J.a.: Evolutionary rates and hbv: issues of rate estimation with bayesian molecular methods. Antiviral therapy (2013)
9. Drummond, A.J., Bouckaert, R.R.: Bayesian Evolutionary Analysis with BEAST. Cambridge University Press, Cambridge (2015)
10. Wu, C.-H., Suchard, M.A., Drummond, A.J.: Bayesian selection of nucleotide substitution models and their site assignments. Mol Biol Evol **30**(3), 669–88 (2013). doi:10.1093/molbev/mss258
11. Bouckaert, R.R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M.A., Rambaut, A., Drummond, A.J.: BEAST 2: a software platform for bayesian evolutionary analysis. PLoS Comput Biol **10**(4), 1003537 (2014). doi:10.1371/journal.pcbi.1003537
12. Tavaré, S.: Some probabilistic and statistical problems in the analysis of dna sequences. Lectures on mathematics in the life sciences **17**, 57–86 (1986)
13. Hasegawa, M., Kishino, H., Yano, T.: Dating the human-ape splitting by a molecular clock of mitochondrial dna. Journal of Molecular Evolution **22**, 160–174 (1985)
14. Pereira, L., Freitas, F., Fernandes, V., Pereira, J.B., Costa, M.D., Costa, S., Máximo, V., Macaulay, V., Rocha, R., Samuels, D.C.: The diversity present in 5140 human mitochondrial genomes. Am J Hum Genet **84**(5), 628–40 (2009). doi:10.1016/j.ajhg.2009.04.013
15. Rosenberg, N.A.: The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly and polyphyly in a coalescent model. Evolution (2003)
16. Jukes, T., Cantor, C.: Evolution of protein molecules. In: Munro, H.N. (ed.) Mammaliam Protein Metabolism, pp. 21–132. Academic Press, New York (1969)
17. Felsenstein, J.: Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of Molecular Evolution **17**, 368–376 (1981)
18. Tamura, K., Nei, M.: Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Molecular Biology and Evolution **10**, 512–526 (1993)
19. Kimura, M.: Estimation of evolutionary distances between homologous nucleotide sequences. Proceedings of the National Academy of Sciences **78**(1), 454–458 (1981)
20. Posada, D.: Using MODELTEST and PAUP* to select a model of nucleotide substitution. Current protocols in bioinformatics, 6–5 (2003)
21. Green, P.J.: Reversible jump markov chain monte carlo computation and bayesian model determination. Biometrika **82**, 711–732 (1995)
22. Pagel, M., Meade, A.: Bayesian analysis of correlated evolution of discrete characters by reversible-jump markov chain monte carlo. The American Naturalist **167**(6), 808–825 (2006)
23. Bouckaert, R.R.: BEASTShell – scripting for bayesian hierarchical clustering. Submitted (2015)
24. Dawid, A.P.: The well-calibrated bayesian. Journal of the American Statistical Association **77**(379), 605–610 (1982)
25. Hayasaka, K., Gojobori, T., Horai, S.: Molecular phylogeny and evolution of primate mitochondrial dna. Molecular Biology and Evolution **5**(6), 626–644 (1988)
26. Gray, R.R., Parker, J., Lemey, P., Salemi, M., Katzourakis, A., Pybus, O.G.: The mode and tempo of hepatitis c virus evolution within and among hosts. BMC evolutionary biology **11**(1), 131 (2011)
27. Heled, J., Drummond, A.J.: Bayesian inference of species trees from multilocus data. Mol Biol Evol **27**(3), 570–80 (2010). doi:10.1093/molbev/msp274
28. Ogilvie, H.A., Heled, J., Xie, D., Drummond, A.J.: Computational performance and statistical accuracy of *beast and comparisons with other methods. Syst Biol **65**(3), 381–96 (2016). doi:10.1093/sysbio/syv118
29. Gavryushkina, A., Welch, D., Stadler, T., Drummond, A.J.: Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. PLoS Comput Biol **10**(12), 1003919 (2014). doi:10.1371/journal.pcbi.1003919
30. Heled, J., Drummond, A.J.: Bayesian inference of population size history from multiple loci. BMC Evol Biol **8**, 289 (2008). doi:10.1186/1471-2148-8-289
31. Stadler, T., Kühnert, D., Bonhoeffer, S., Drummond, A.J.: Birth-death skyline plot reveals temporal changes of epidemic spread in hiv and hepatitis c virus (hcv). Proc Natl Acad Sci U S A **110**(1), 228–33 (2013). doi:10.1073/pnas.1207965110

Appendix

list of all transition/tranversion split models

model number	r_{ac}	r_{ag}	r_{at}	r_{cg}	r_{ct}	r_{gt}	name
0	1	1	1	1	1	1	JC69/F81
1	1	2	1	1	2	1	K80/HKY
2	1	2	1	1	2	3	
3	1	2	1	1	3	1	TN93
4	1	2	1	1	3	4	
5	1	2	1	3	2	1	
6	1	2	1	3	2	3	
7	1	2	1	3	2	4	
8	1	2	1	3	4	1	
9	1	2	1	3	4	3	
10	1	2	1	3	4	5	
11	1	2	3	1	2	1	
12	1	2	3	1	2	3	
13	1	2	3	1	2	4	
14	1	2	3	1	4	1	
15	1	2	3	1	4	3	
16	1	2	3	1	4	5	
17	1	2	3	3	2	1	K81
18	1	2	3	3	2	3	
19	1	2	3	3	2	4	
20	1	2	3	3	4	1	TIM
21	1	2	3	3	4	3	
22	1	2	3	3	4	5	
23	1	2	3	4	2	1	
24	1	2	3	4	2	3	
25	1	2	3	4	2	4	
26	1	2	3	4	2	5	TVM
27	1	2	3	4	5	1	
28	1	2	3	4	5	3	
29	1	2	3	4	5	4	
30	1	2	3	4	5	6	GTR

List of all named models, and potential models to split into

JC69/F81 111111 : 121121

K80/HKY 121121 : 121131, 123321

TN93 121131 : 123341

K81 123321 : 123341, 123421

TIM 123341 : 123345, 123451

TVM 123421 : 123451

123345 : 123456

123451 : 123456

GTR 123456 :

List of all transition/tranversionsplit models, and potential models to split into

111111 : 121121
121121 : 121123, 121131, 121321, 121323, 123121, 123123, 123321, 123323
121123 : 121134, 121324, 123124, 123324
121131 : 121134, 121341, 121343, 123141, 123143, 123341, 123343
121134 : 121345, 123145, 123345
121321 : 121324, 121341, 123421, 123423
121323 : 121324, 121343, 123424
121324 : 121345, 123425
121341 : 121345, 123451, 123453
121343 : 121345, 123454
121345 : 123456
123121 : 123124, 123141, 123421, 123424
123123 : 123124, 123143, 123423
123124 : 123145, 123425
123141 : 123145, 123451, 123454
123143 : 123145, 123453
123145 : 123456
123321 : 123324, 123341, 123421
123323 : 123324, 123343, 123423, 123424
123324 : 123345, 123425
123341 : 123345, 123451
123343 : 123345, 123453, 123454
123345 : 123456
123421 : 123425, 123451
123423 : 123425, 123453
123424 : 123425, 123454
123425 : 123456
123451 : 123456
123453 : 123456
123454 : 123456
123456 :

List of all reversible models, and potential models to split into

111111 : 111112, 111121, 111122, 111211, 111212, 111221, 111222, 112111, 112112, 112121, 112122, 112211, 112212, 112221, 112222, 121111, 121112, 121121, 121122, 121211, 121212, 121221, 121222, 122111, 122112, 122121, 122122, 122211, 122212, 122221, 122222
111112 : 111123, 111213, 111223, 112113, 112123, 112213, 112223, 121113, 121123, 121213, 121223, 122113, 122123, 122213, 122223
111121 : 111123, 111231, 111232, 112131, 112132, 112231, 112232, 121131, 121132, 121231, 121232, 122131, 122132, 122231, 122232
111122 : 111123, 111233, 112133, 112233, 121133, 121233, 122133, 122233
111123 : 111234, 112134, 112234, 121134, 121234, 122134, 122234
111211 : 111213, 111231, 111233, 112311, 112312, 112321, 112322, 121311, 121312,

121321, 121322, 122311, 122312, 122321, 122322
111212 : 111213, 111232, 112313, 112323, 121313, 121323, 122313, 122323
111213 : 111234, 112314, 112324, 121314, 121324, 122314, 122324
111221 : 111223, 111231, 112331, 112332, 121331, 121332, 122331, 122332
111222 : 111223, 111232, 111233, 112333, 121333, 122333
111223 : 111234, 112334, 121334, 122334
111231 : 111234, 112341, 112342, 121341, 121342, 122341, 122342
111232 : 111234, 112343, 121343, 122343
111233 : 111234, 112344, 121344, 122344
111234 : 112345, 121345, 122345
112111 : 112113, 112131, 112133, 112311, 112313, 112331, 112333, 123111, 123112,
123121, 123122, 123211, 123212, 123221, 123222
112112 : 112113, 112132, 112312, 112332, 123113, 123123, 123213, 123223
112113 : 112134, 112314, 112334, 123114, 123124, 123214, 123224
112121 : 112123, 112131, 112321, 112323, 123131, 123132, 123231, 123232
112122 : 112123, 112132, 112133, 112322, 123133, 123233
112123 : 112134, 112324, 123134, 123234
112131 : 112134, 112341, 112343, 123141, 123142, 123241, 123242
112132 : 112134, 112342, 123143, 123243
112133 : 112134, 112344, 123144, 123244
112134 : 112345, 123145, 123245
112211 : 112213, 112231, 112233, 112311, 123311, 123312, 123321, 123322
112212 : 112213, 112232, 112312, 112313, 123313, 123323
112213 : 112234, 112314, 123314, 123324
112221 : 112223, 112231, 112321, 112331, 123331, 123332
112222 : 112223, 112232, 112233, 112322, 112323, 112332, 112333, 123333
112223 : 112234, 112324, 112334, 123334
112231 : 112234, 112341, 123341, 123342
112232 : 112234, 112342, 112343, 123343
112233 : 112234, 112344, 123344
112234 : 112345, 123345
112311 : 112314, 112341, 112344, 123411, 123412, 123421, 123422
112312 : 112314, 112342, 123413, 123423
112313 : 112314, 112343, 123414, 123424
112314 : 112345, 123415, 123425
112321 : 112324, 112341, 123431, 123432
112322 : 112324, 112342, 112344, 123433
112323 : 112324, 112343, 123434
112324 : 112345, 123435
112331 : 112334, 112341, 123441, 123442
112332 : 112334, 112342, 123443
112333 : 112334, 112343, 112344, 123444
112334 : 112345, 123445
112341 : 112345, 123451, 123452
112342 : 112345, 123453
112343 : 112345, 123454

112344 : 112345, 123455
112345 : 123456
121111 : 121113, 121131, 121133, 121311, 121313, 121331, 121333, 123111, 123113, 123131, 123133, 123311, 123313, 123331, 123333
121112 : 121113, 121132, 121312, 121332, 123112, 123132, 123312, 123332
121113 : 121134, 121314, 121334, 123114, 123134, 123314, 123334
121121 : 121123, 121131, 121321, 121323, 123121, 123123, 123321, 123323
121122 : 121123, 121132, 121133, 121322, 123122, 123322
121123 : 121134, 121324, 123124, 123324
121131 : 121134, 121341, 121343, 123141, 123143, 123341, 123343
121132 : 121134, 121342, 123142, 123342
121133 : 121134, 121344, 123144, 123344
121134 : 121345, 123145, 123345
121211 : 121213, 121231, 121233, 121311, 123211, 123213, 123231, 123233
121212 : 121213, 121232, 121312, 121313, 123212, 123232
121213 : 121234, 121314, 123214, 123234
121221 : 121223, 121231, 121321, 121331, 123221, 123223
121222 : 121223, 121232, 121233, 121322, 121323, 121332, 121333, 123222
121223 : 121234, 121324, 121334, 123224
121231 : 121234, 121341, 123241, 123243
121232 : 121234, 121342, 121343, 123242
121233 : 121234, 121344, 123244
121234 : 121345, 123245
121311 : 121314, 121341, 121344, 123411, 123413, 123431, 123433
121312 : 121314, 121342, 123412, 123432
121313 : 121314, 121343, 123414, 123434
121314 : 121345, 123415, 123435
121321 : 121324, 121341, 123421, 123423
121322 : 121324, 121342, 121344, 123422
121323 : 121324, 121343, 123424
121324 : 121345, 123425
121331 : 121334, 121341, 123441, 123443
121332 : 121334, 121342, 123442
121333 : 121334, 121343, 121344, 123444
121334 : 121345, 123445
121341 : 121345, 123451, 123453
121342 : 121345, 123452
121343 : 121345, 123454
121344 : 121345, 123455
121345 : 123456
122111 : 122113, 122131, 122133, 122311, 122313, 122331, 122333, 123111
122112 : 122113, 122132, 122312, 122332, 123112, 123113
122113 : 122134, 122314, 122334, 123114
122121 : 122123, 122131, 122321, 122323, 123121, 123131
122122 : 122123, 122132, 122133, 122322, 123122, 123123, 123132, 123133
122123 : 122134, 122324, 123124, 123134

122131 : 122134, 122341, 122343, 123141
122132 : 122134, 122342, 123142, 123143
122133 : 122134, 122344, 123144
122134 : 122345, 123145
122211 : 122213, 122231, 122233, 122311, 123211, 123311
122212 : 122213, 122232, 122312, 122313, 123212, 123213, 123312, 123313
122213 : 122234, 122314, 123214, 123314
122221 : 122223, 122231, 122321, 122331, 123221, 123231, 123321, 123331
122222 : 122223, 122232, 122233, 122322, 122323, 122332, 122333, 123222, 123223, 123232, 123233, 123322, 123323, 123332, 123333
122223 : 122234, 122324, 122334, 123224, 123234, 123324, 123334
122231 : 122234, 122341, 123241, 123341
122232 : 122234, 122342, 122343, 123242, 123243, 123342, 123343
122233 : 122234, 122344, 123244, 123344
122234 : 122345, 123245, 123345
122311 : 122314, 122341, 122344, 123411
122312 : 122314, 122342, 123412, 123413
122313 : 122314, 122343, 123414
122314 : 122345, 123415
122321 : 122324, 122341, 123421, 123431
122322 : 122324, 122342, 122344, 123422, 123423, 123432, 123433
122323 : 122324, 122343, 123424, 123434
122324 : 122345, 123425, 123435
122331 : 122334, 122341, 123441
122332 : 122334, 122342, 123442, 123443
122333 : 122334, 122343, 122344, 123444
122334 : 122345, 123445
122341 : 122345, 123451
122342 : 122345, 123452, 123453
122343 : 122345, 123454
122344 : 122345, 123455
122345 : 123456
123111 : 123114, 123141, 123144, 123411, 123414, 123441, 123444
123112 : 123114, 123142, 123412, 123442
123113 : 123114, 123143, 123413, 123443
123114 : 123145, 123415, 123445
123121 : 123124, 123141, 123421, 123424
123122 : 123124, 123142, 123144, 123422
123123 : 123124, 123143, 123423
123124 : 123145, 123425
123131 : 123134, 123141, 123431, 123434
123132 : 123134, 123142, 123432
123133 : 123134, 123143, 123144, 123433
123134 : 123145, 123435
123141 : 123145, 123451, 123454
123142 : 123145, 123452

123143 : 123145, 123453
123144 : 123145, 123455
123145 : 123456
123211 : 123214, 123241, 123244, 123411
123212 : 123214, 123242, 123412, 123414
123213 : 123214, 123243, 123413
123214 : 123245, 123415
123221 : 123224, 123241, 123421, 123441
123222 : 123224, 123242, 123244, 123422, 123424, 123442, 123444
123223 : 123224, 123243, 123423, 123443
123224 : 123245, 123425, 123445
123231 : 123234, 123241, 123431
123232 : 123234, 123242, 123432, 123434
123233 : 123234, 123243, 123244, 123433
123234 : 123245, 123435
123241 : 123245, 123451
123242 : 123245, 123452, 123454
123243 : 123245, 123453
123244 : 123245, 123455
123245 : 123456
123311 : 123314, 123341, 123344, 123411
123312 : 123314, 123342, 123412
123313 : 123314, 123343, 123413, 123414
123314 : 123345, 123415
123321 : 123324, 123341, 123421
123322 : 123324, 123342, 123344, 123422
123323 : 123324, 123343, 123423, 123424
123324 : 123345, 123425
123331 : 123334, 123341, 123431, 123441
123332 : 123334, 123342, 123432, 123442
123333 : 123334, 123343, 123344, 123433, 123434, 123443, 123444
123334 : 123345, 123435, 123445
123341 : 123345, 123451
123342 : 123345, 123452
123343 : 123345, 123453, 123454
123344 : 123345, 123455
123345 : 123456
123411 : 123415, 123451, 123455
123412 : 123415, 123452
123413 : 123415, 123453
123414 : 123415, 123454
123415 : 123456
123421 : 123425, 123451
123422 : 123425, 123452, 123455
123423 : 123425, 123453
123424 : 123425, 123454

123425 : 123456
123431 : 123435, 123451
123432 : 123435, 123452
123433 : 123435, 123453, 123455
123434 : 123435, 123454
123435 : 123456
123441 : 123445, 123451
123442 : 123445, 123452
123443 : 123445, 123453
123444 : 123445, 123454, 123455
123445 : 123456
123451 : 123456
123452 : 123456
123453 : 123456
123454 : 123456
123455 : 123456
123456 :

List of parameters reported by the model

- BMT_ModelInidicator is the index of the substitution model as listed in the Appendix.
- substmodel is the model represented as a 6-digit number, where the position of the digit refers to rates ac, ag, at, cg, ct and gt respectively, and equal digits indicates that rates are shared, so 111111 is Jukes Cantor (if frequencies are kept equal), 121121 is HKY, 123456 is GTR etc.
- rateAC,. . . ,rateGT are the rates according to the model. ESSs should be good for these rates, but if you plot joint-marginals of pairs you may find high correlation between some of these rates.
- BMT_Rates.1 to 6 are the rates used to build up the rate matrix. If only low parameter models are samples, the higher rates will be sampled very infrequently, and you should expect low ESSs for them. Correlation between pairs of rates should be low.
- BM_gammaShape is the gamma shape parameter as it is being sampled. For parts of the chain that gamma rate heterogeneity is switched off, the parameter will not be sampled, and the trace will show periods where the parameter is stuck.
- hasGammaRates indicates whether gamma rate heterogeneity it used (1) or not used (0). The mean can be interpreted as the proportion of time that gamma rate heterogeneity is switched on.
- ActiveGammaShape is the gamma shape parameter when it is sampled, but it is zero when it it not sampled. To get the estimate of the mean of the shape parameter, divide the mean ActiveGammaShape by the mean of hasGammaRates.
- BMT_ProportionInvariable, hasInvariableSites and ActivePropInvariable are the value for proportion invariable similar to BMG_gammaShape, hasGammaRates and ActiveGammaShape respectively.

- hasEqualFreqs indicates whether equal frequencies are used and the mean can be interpreted as the proportion of time that equal frequencies is used. When empirical frequencies are used, this parameter is not reported.

95% HPD of models for Primates data

empirical freqs			estimated freqs		
posterior support	cumulative support	model	posterior support	cumulative support	model
21.12 %	21.12 %	121323	16.19%	16.19%	121121
17.48 %	38.60 %	123424	14.09%	30.27%	121123
13.71 %	52.31 %	123324	10.25%	40.53%	121323
10.13 %	62.44 %	123323	9.71%	50.24%	121131
7.82 %	70.26 %	121324	5.31%	55.55%	121134
7.74 %	78.00 %	121123	4.63%	60.18%	123323
5.48 %	83.48 %	123425	4.28%	64.46%	121321
2.87 %	86.35 %	123423	3.94%	68.40%	121324
2.69 %	89.03 %	121343	2.98%	71.38%	121343
2.24 %	91.28 %	123454	2.79%	74.17%	123324
1.50 %	92.78 %	123345	2.43%	76.60%	121341
1.38 %	94.16 %	123343	2.37%	78.97%	123123
1.27 %	95.42 %	123124	2.24%	81.21%	123343
			2.09%	83.30%	123424
			1.96%	85.26%	123321
			1.96%	87.21%	123124
			1.94%	89.16%	123121
			1.52%	90.68%	123143
			1.36%	92.03%	123341
			1.32%	93.36%	123423
			1.17%	94.52%	123141
			1.14%	95.67%	121345