Bouckaert and Drummond *BMC Evolutionary Biology* (2017) 17:42
DOI 10.1186/s12862-017-0890-6

BMC Evolutionary Biology

SOFTWARE Open Access

bModelTest: Bayesian phylogenetic site model averaging and model comparison

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

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 probability 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.

Co-estimation of the substitution model for a nucleotide alignment can be achieved by sampling all possible reversible models [7], or just a nested set of models [8], using either reversible jump MCMC or stochastic Bayesian variable selection [9]. The CAT-GTR model [10, 11] solves a related problem by providing a mixture model over sites that often fits better than using any single

*Correspondence: remco@cs.auckland.ac.nz
[1]Centre for Computational Evolution, University of Auckland, Auckland, New Zealand
[2]Department of Computer Science, University of Auckland, Auckland, New Zealand
Full list of author information is available at the end of the article

model for all sites. Wu et al. [12] use reversible jump for both substitution models and partitions and furthermore sample the use of gamma rate heterogeneity for each site category. However, since the method divides sites among a set of substitution models, it does not address invariable sites, and only considers a limited set of five (K80, F81, HKY85, TN93, and GTR) substitution models.

Here 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 [13] 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.

Implementation

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 123,456 corresponds to a model where all rates are independent, named the general time reversible (GTR) model [14]. Model 121121 corresponds to the HKY model [15] 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 646,646 and 212,212 represent HKY, we only use 121,121.

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 transversions (the other substitutions) [16, 17]. 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 111,111, where all rates are grouped. This reduces the 203 models to 31 models (see Fig. 1 and details in Additional file 1: Appendix). Alternatively, if one is interested in using named models, we can restrict further to include only Jukes Cantor [18, 19] (111,111), HKY [15] (121,121), TN93 [20] (121,131), K81 [21] (123,321), TIM [22] (123,341), TVM [22] (123,425),and GTR [14] (123,456). However, to facilitate stepping between TIM and GTR during the MCMC (see proposals below) we like to use nested models, and models 123,345 and 123,324 provide intermediates between TIM and GTR, as well as K81 and TVM, leaving us with a set of 9 models (Fig. 1).

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 [16, 17], 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

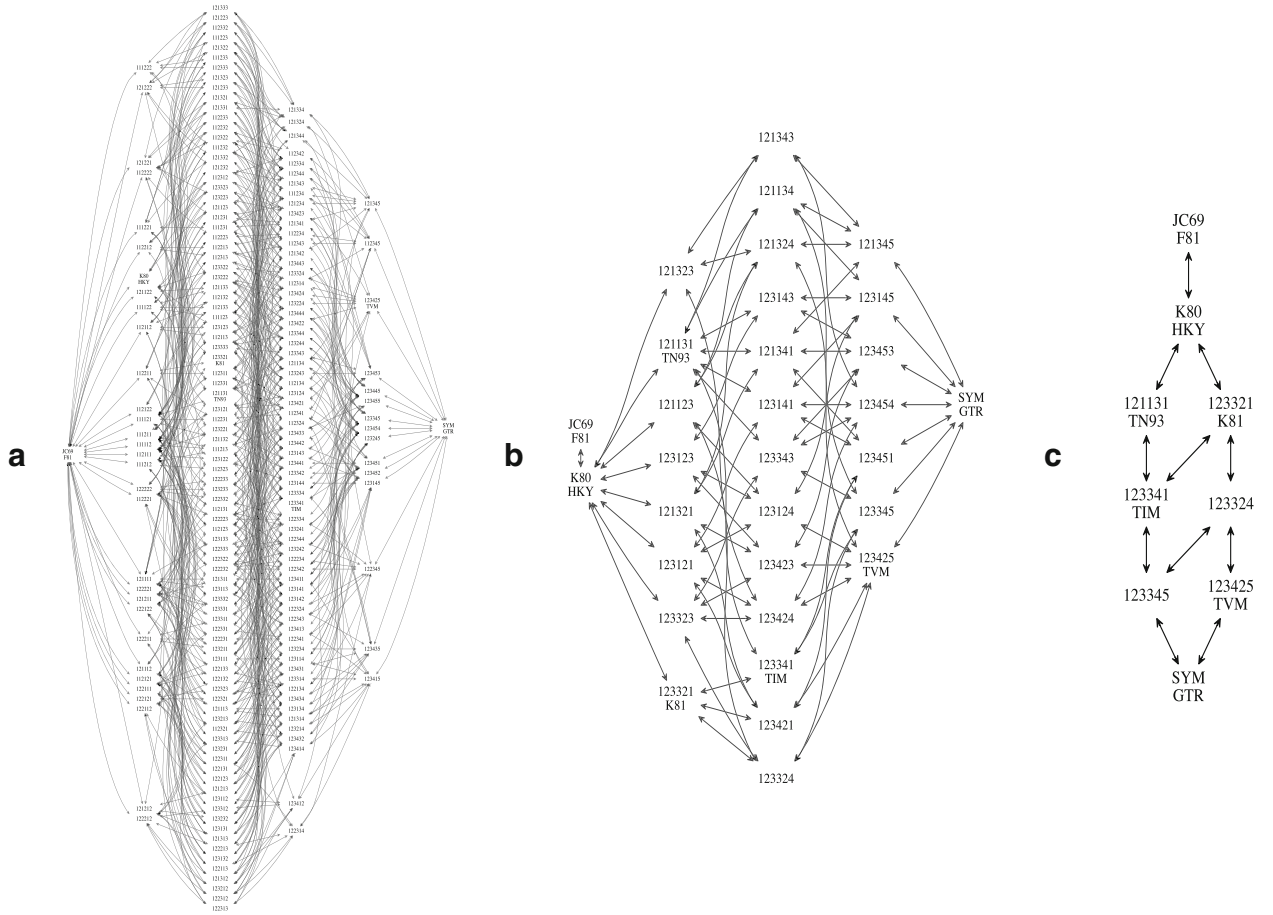

Fig. 1 Model spaces. The model spaces supported by bModelTest. **a** All reversible models, **b** transition/transversion split models, and **c** named models. *Arrows* indicate which models can be reached by splitting a model. Note all models with the same number of groupings are at the same height

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 Fig. 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 [23] 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 [23].

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 Fig. 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. For any set of substitution models organised in an adjacency graph our merge/split operator applies, making our graph-based method easier to generalise to other model sets (e.g. the one used in [24]). If there are no candidates to split (that is, model $M = 123,456$ is GTR) the proposal returns the current state (this proposal is important to guarantee uniform sampling of models). Likewise, when attempting to merge model $M = 111,111$, the current state is proposed ($M' = 111,111$). 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

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

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

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

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 [13], 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 transversions 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 (Bouckaert, RR: BEASTShell – scripting for bayesian hierarchical clustering, submitted) 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 against 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.

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. 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

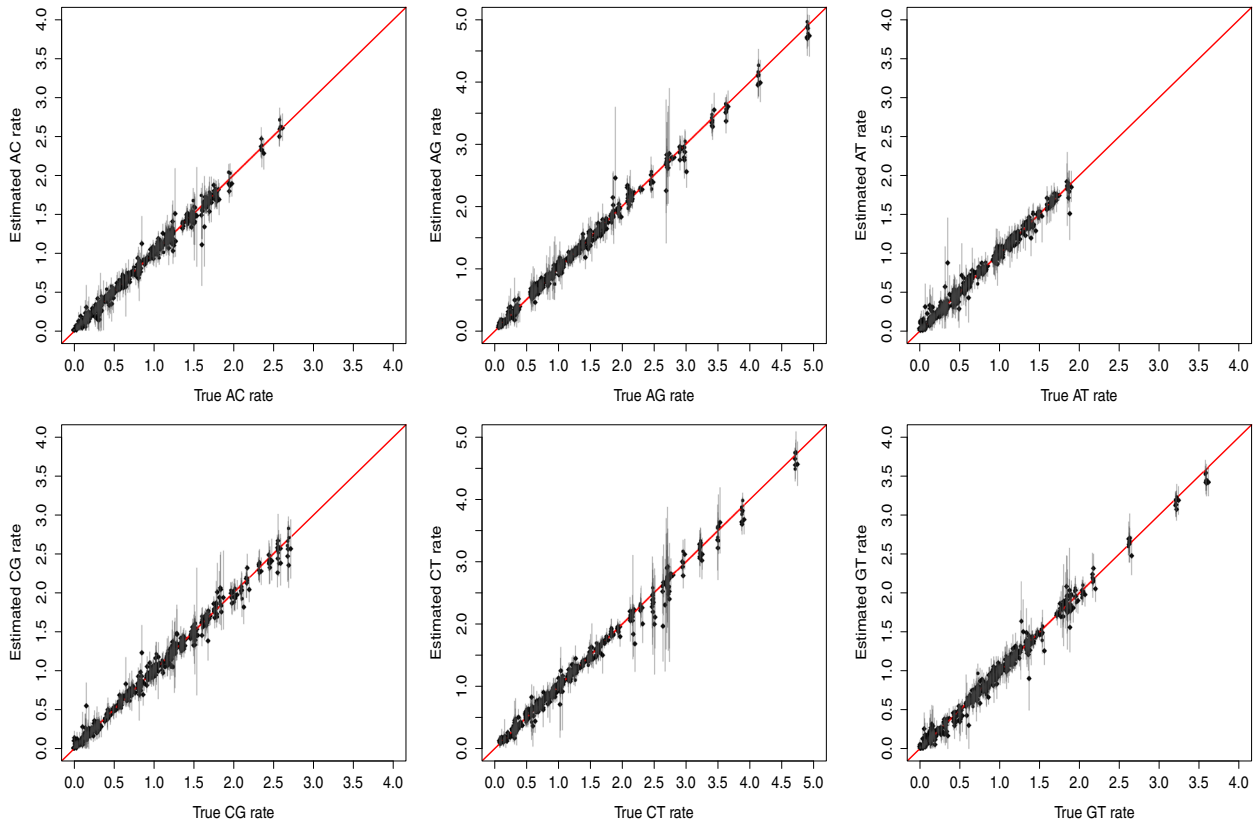

Fig. 2 Accuracy of estimated substitution rates. True rates (*horizontal*) against estimated rates (*vertical*) in simulated data for 3 taxa. In reading order, rate AC, AG, AT, CG, CT and GT. *Diamonds* are for estimates when no rate heterogeneity was used to simulate the data, *circles* are for estimates with rate heterogeneity. *Error bars* represent 95% HPD intervals for each estimate

equal frequencies, the latter four with unequal frequencies. The last row shows results averaged over all 800 experiments. On average, one would expect the coverage to be 95% if simulations are drawn from the prior [25], so each entry in Table 1 has an expected value of 95, but can deviate due to small sample size. According to the binomial probability distribution there is a $\sim$ 1.1% chance of seeing 89 or less successes when sampling 100 times with a success rate of 0.95. The sample size for the mean rows is 800, so is expected to be much closer to 95%.

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.

To investigate robustness of the approach, we repeated the study with a log normal uncorrelated relaxed clock [26] with a gamma($\alpha = 30, \beta = 0.005$) prior over the

Table 1 Coverage summary for simulation study

	Site	Rate coverage						Mean	Subst. Model
Freqs	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

	Site	Site model			Frequency	Frequency coverage			
Freqs	model	coverage	α	p_{inv}	coverage	A	C	G	T
Equal	plain	100			100	100	100	100	100
Equal	+G	96	94		100	100	100	100	100
Equal	+I	98		95	100	100	100	100	100
Equal	+G+I	99	89	88	100	100	100	100	100
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	97.38	97.88	97.13	97.38

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*

standard deviation for the log normal distribution. Trees with 5 taxa were randomly sampled from a Yule prior with log normal distribution (the birth rate was drawn from a distribution with a mean of the rate of 5.5, and a standard-deviation of the log-rate of 0.048) giving trees with mean height $\approx$ 0.25 and 95% HPD interval of 0.015 to 0.7. The study as outlined above was repeated, and results are summarised in Additional file 1: Table S1, which looks very similar to that of Table 1. So, we conclude that the model is not sensitive to small variation in molecular clock rates among branches.

Figure 3 shows histograms of estimated posterior probability of gamma-distributed rate heterogeneity across sites for the data sets simulated over 5 taxa. When data was generated without gamma-distributed rate heterogeneity across sites, the posterior probability was often estimated to be close to zero (left of Fig. 3), while the posterior probability was estimated to be close to one for most of the analyses on data in which gamma rate heterogeneity was present (middle of Fig. 3).[1] When rate heterogeneity was present, shape estimates were fairly close to the ones used to generate the data (right of Fig. 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 Fig. 3 but for the proportion of invariable sites for 5 taxa. [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

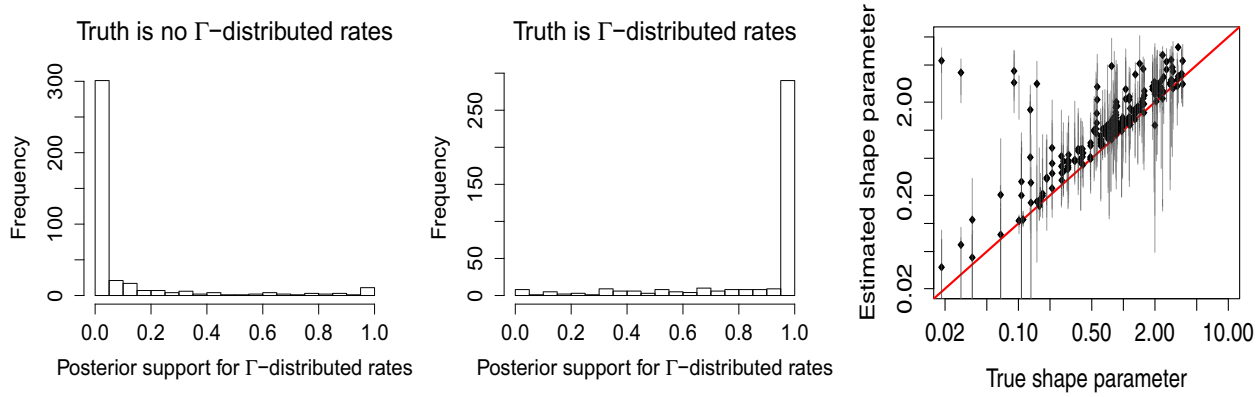

Fig. 3 Accuracy of inference of rate heterogeneity across sites. Posterior probability for inclusion of gamma rate heterogeneity when the data is generated without (*left*) and with (*middle*) rate heterogeneity for 5 taxa. Right, True gamma shape parameter (*horizontal*) against estimated shape parameter (*vertical*) when rate heterogeneity is used to generate the data

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

The same study with 5 taxa was repeated with the substitution model fixed to HKY and GTR, but estimating the other parts of the model. Results are summarised in Additional file 1: Tables S2 and S3 respectively. Fixing the model to HKY results in severe degradation of accuracy in all parameter and model estimates. The lack of coverage of frequency estimates when the true model has equal frequencies suggests that lack of degrees of freedom in substitution model parameters is compensated by estimating frequencies instead of keeping them equal. So substitution model misspecification can result in considerable misspecification of the remainder of the model. Results when fixing the substitution model to GTR shows a table with results very similar to that of bModelTest, however the substitution model parameters have on average a 95% HPD interval of size 0.17 while that of bModelTest is only 0.13. The extra parameters that need to be estimated for GTR compared to bModelTest result in more uncertain estimates, and thus more uncertainty in the analysis.

To see the impact of the model set, the experiment was repeated with sampling from all 203 reversible models instead of using only the 30 transition/transversion split models. Results are shown in Additional file 1: Table S4, which do not differ substantially from Table 1. Further, to investigate the effect of the number of taxa and sequence length, the study was repeated with 16 taxa and sequence lengths 1K and 0.5K base pairs long under a relaxed clock as before. Results are summarised in Additional file 1: Tables S5 and S6 respectively. The tables do not show significant differences to Table 1 or

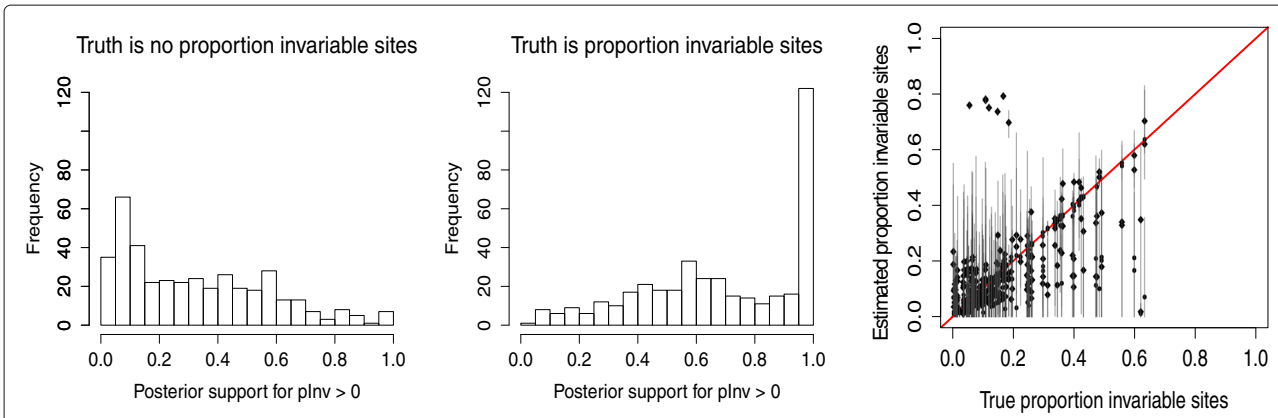

Fig. 4 Accuracy of inference of proportion of invariant sites. Posterior probability for inclusion of a proportion of invariant sites when the data is generated without (*left*) and with (*middle*) invariant sites for 5 taxa. Right, empirical proportion invariant in alignment (*horizontal*) against estimated proportion of invariant sites (*vertical*) when a proportion invariable category is used to generate the data

degradation with decreasing sequence length, so the ability of our Bayesian method to correctly estimate the posterior distribution of substitution models and their parameters does not appear to depend substantially on sequence length or number of taxa.

Comparison with jModelTest

We ran jModelTest version 2.1.10 [3] on the sequence data used for the last simulation study with 5 taxa (using all reversible models, since only that set is the same for both jModelTest and bModelTest) and the two simulation studies using 16 taxa and compared the substitution model coverage (with settings -BIC -AIC -f -g 4 -i -s 203). For each dataset, we collected the top models according to the AIC and BIC criteria such that the cumulative weight exceeded 95% of the models as shown in the jModelTest output and registered whether the true model was contained in the resulting set. Results are summarised in Additional file 1: Table S7, which shows that both AIC and BIC do not cover the true model 95% of the time as would be desirable. For some combinations the coverage is close to the desirable value (89.4% for AIC with 5 taxa) and for some it is much lower (61.1% for BIC with 0.5K length sequences and 16 taxa). Coverage of both AIC and BIC appears to decrease with increasing number of taxa and decreasing sequence length, although we have not attempted a systematic study. In contrast bModelTest has a coverage of $\sim$ 95% for all scenarios. jModelTest uses a single maximum likelihood tree and it seems that increasing uncertainty in the true tree (by increasing the number of taxa or decreasing sequence length) results in an increasing chance of incorrect model weights from jModelTest. For BIC, we find substantially less coverage of jModelTest than the around 90% model coverage reported in a previous study [3]. This is probably because our data contains a larger amount of uncertainty due to shorter sequences and tree lengths. Another factor is that we use different priors. For example, we use a Beta(1,4) for the proportion of invariable sites, while the previous study [3] used a Beta(1,3) that was then truncated to the interval [0.2,0.8], thereby avoiding extreme values which might cause difficulties. To confirm this we produced simulated data more closely matched to previously published experiments (with 40 taxa, sequences of 2500 base pairs, models selected uniformly from the 11 named models, tree length with mean of 6.5, truncated prior for invariable sites, BIC criterion) and obtained a coverage of 93.8% for the 95% credible set and 89.5% coverage by the best fitting model, similar to the results in [3].

In practice, users of Bayesian phylogenetic packages only use the most highly weighted model returned by jModelTest. Additional file 1: Table S7 shows how often the best fitting model according to AIC and BIC matches the true model, which ranges from 73.9% for BIC on 5 taxa to 30.8% for AIC on 0.5K length sequences and 16 taxa, suggesting that the probability of model misspecification using this approach increases with phylogenetic uncertainty.

To compare the application of bModelTest to jModelTest (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 [27] (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 (=121,323) has the highest posterior probability using bModelTest (21.12% see Additional file 1: Appendix for full list of supported models) when empirical frequencies are used. However, when frequencies are allowed to be estimated, HKY has highest posterior probability (16.19%), while TPM2 (10.25%) has less posterior probability then model 121,123 (14.09%). So, using a heuristic maximum likelihood approach (jModelTest and/or empirical frequencies) makes a difference in the substitution model being preferred. Figure 5 left shows the posterior probabilities 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 Fig. 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 [28]. 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 preferred for the first partition have transition parameters split while for the second partition models where partitions are shared have higher posterior probability, resulting in quite distinct model coverage images. For the first

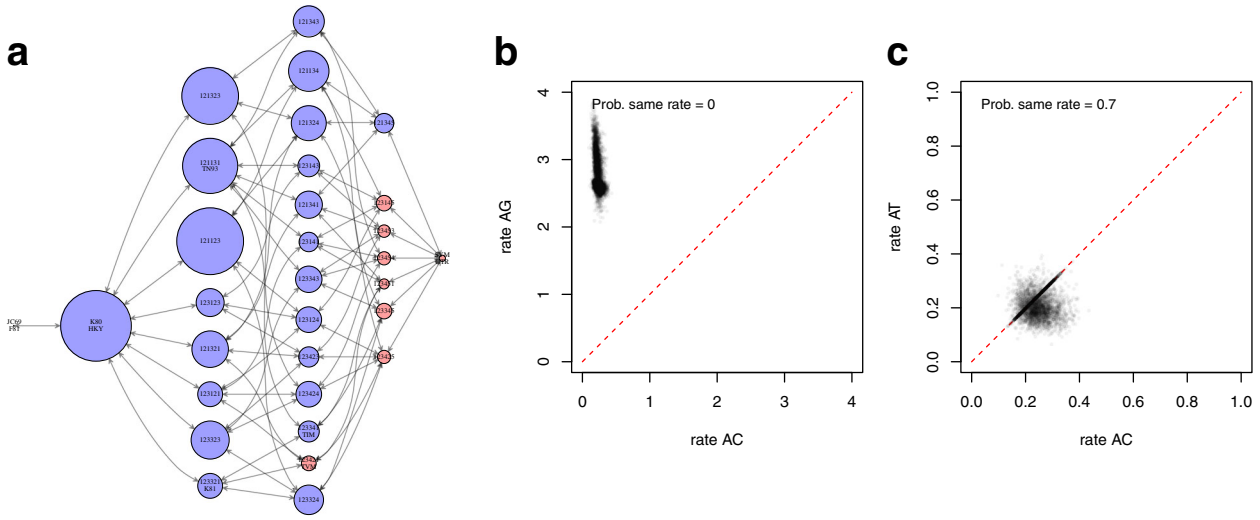

Fig. 5 Posterior inference on primate data. Model distribution for primate data using the transition/transversion split models (*left*). Numbers on x-axis correspond to models in Additional file 1: Appendix. The *middle panel* plots rates $A \leftrightarrow C$ versus $A \leftrightarrow G$ (*middle*) and the *right panel* plots $A \leftrightarrow C$ versus $A \leftrightarrow T$

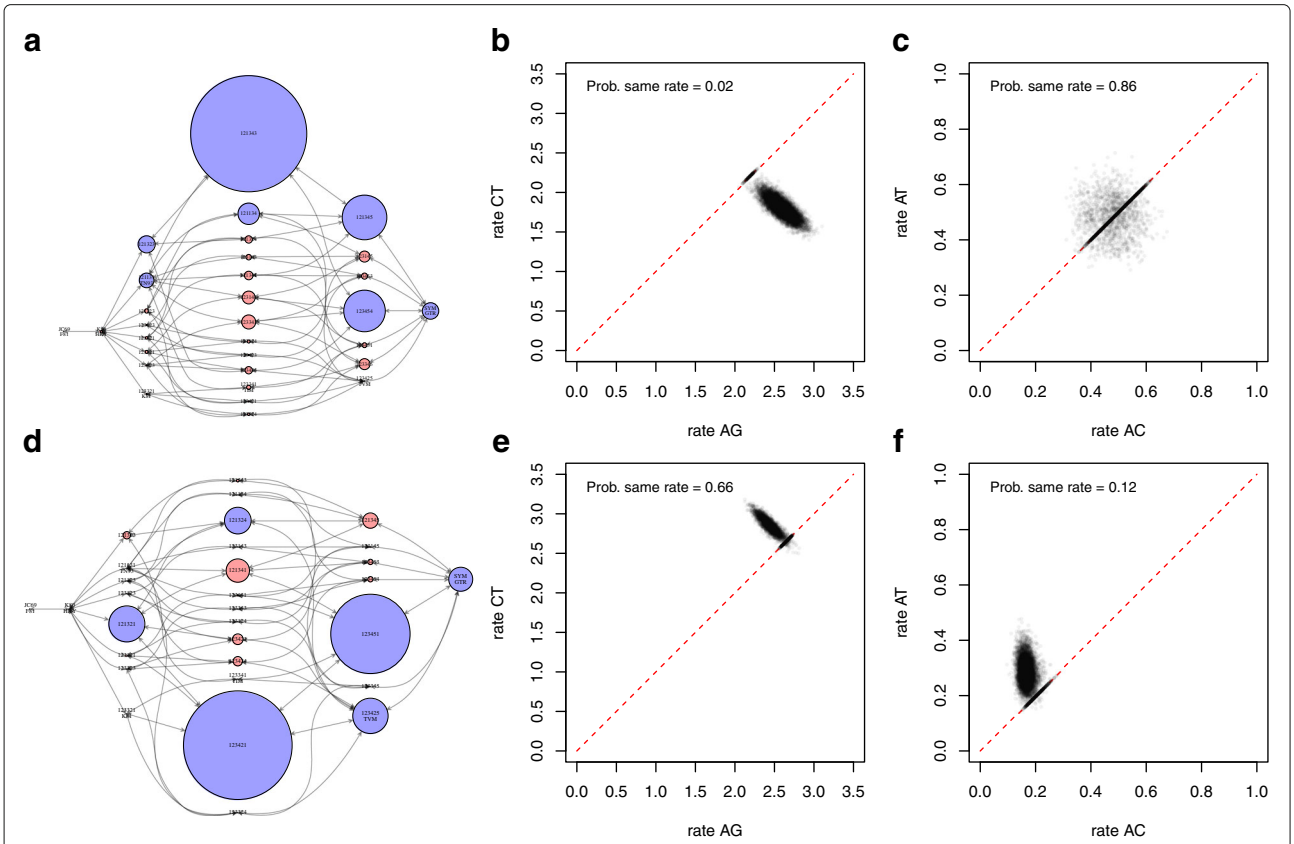

Fig. 6 Posterior inference on HCV data. Like Fig. 5, but the data is split into two partitions, the first containing codon positions 1+2 (panel **a**, **b** and **c**) and second containing codon position 3 (panel **d**, **e** and **f**). The partitions support distinctly different site models. The *left 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 transversion rates $A \leftrightarrow C$ versus $A \leftrightarrow T$

partition, jModelTest recommends TIM2+I+G. TIM2 is model 121,343, the model with highest posterior probability according to bModelTest, as shown in Fig. 6. For the second partition, jModelTest recommends TVM+G, and though TVM is in the 95% credible set, it has a lower posterior probability than model 123421, which gets the highest posterior probability according to bModelTest. Running jModelTest on all 203 models, model 123,451 is preferred by both AIC and BIC, even though 123421 was considered by jModelTest. Again, we see a difference in heuristic likelihood and full Bayesian approaches. The correlation between transition rates $A \leftrightarrow G$ and $C \leftrightarrow T$ as well as between two transversion rates $A \leftrightarrow C$ and $A \leftrightarrow T$ are shown in Fig. 6 top middle and right for the first partition and Fig. 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 posterior probability for 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 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 [29, 30], various forms of phylogeographical analyses, sampled ancestor analysis [31], demographic reconstruction using coalescent [32], birth death skyline analysis [33], *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.

A promising direction for further research would be to incorporate efficient averaging over partitioning of the alignment [10–12] to the site model averaging approach described here.

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.

Availability and requirements

Project name: bModelTest
Project home page: https://github.com/BEAST2-Dev/bModelTest/
Operating systems: Windos, OSX, Linux and any other OS
Programming language: Java
Other requirements: requires BEAST 2 (from http://beast2) Licence: LGPL.

Endnotes

[1] Estimated shape parameters only take values of the shape parameter in account in the portion of the posterior sample where gamma rate heterogeneity indicator is 1.

[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.

Additional file

Additional file 1: Appendix. (PDF 459 kb)

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

Acknowledgements

Not applicable.

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.

Availability of data and material

See supplementary information for material for the simulation study as well as the primates and HCV analyses. The method is implemented in the bModelTest package of BEAST 2 available from http://beast2.org under LGPL license. Source code, installation instructions and documentation can be found at https://github.com/BEAST2-Dev/bModelTest.

Authors' contributions
RRB designed and developed software. AJD and RRB designed experiments. RRB executed experiments. AJD and RRB analyzed the data. AJD and RRB prepared the figures. AJD and RRB wrote the manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

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

Received: 19 June 2016 Accepted: 19 January 2017
Published online: 06 February 2017

References

1. Posada D, Crandall KA. Modeltest: testing the model of dna substitution. Bioinformatics. 1998;14(9):817–8.
2. Posada D. jModelTest: phylogenetic model averaging. Mol Biol Evol. 2008;25(7):1253–56.
3. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods. 2012;9(8):772–2.
4. Yang Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol. 1994;39(3):306–14. doi:10.1007/BF00160154.
5. Gu X, Fu YX, Li WH. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. Mol Biol Evol. 1995;12(4): 546–7.
6. Waddell P, Penny D. Evolutionary trees of apes and humans from DNA sequences In: Lock AJ, Peters CR, editors. Handbook of Symbolic Evolution. Oxford: Clarendon Press; 1996. p. 53–73.
7. Huelsenbeck JP, Larget B, Alfaro ME. Bayesian phylogenetic model selection using reversible jump markov chain monte carlo. Mol Biol Evol. 2004;21(6):1123–33. doi:10.1093/molbev/msh123.
8. Bouckaert RR, Alvarado-Mora M, Rebello Pinho Ja. Evolutionary rates and hbv: issues of rate estimation with bayesian molecular methods. Antivir Ther. 2013;18(3 Pt B):497–503.
9. Drummond AJ, Bouckaert RR. Bayesian evolutionary analysis with BEAST. Cambridge: Cambridge University Press; 2015.
10. Lartillot N, Lepage T, Blanquart S. Phylobayes 3: a bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics. 2009;25(17):2286–288.
11. Lartillot N, Philippe H. A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol. 2004;21(6):1095–109.
12. Wu CH, Suchard MA, Drummond AJ. Bayesian selection of nucleotide substitution models and their site assignments. Mol Biol Evol. 2013;30(3): 669–88. doi:10.1093/molbev/mss258.
13. Bouckaert RR, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ. BEAST 2: a software platform for bayesian evolutionary analysis. PLoS Comput Biol. 2014;10(4):1003537. doi:10.1371/journal.pcbi.1003537.
14. Tavaré S. Some probabilistic and statistical problems in the analysis of dna sequences. Lect Math Life Sci. 1986;17:57–86.
15. Hasegawa M, Kishino H, Yano T. Dating the human-ape splitting by a molecular clock of mitochondrial dna. J Mol Evol. 1985;22:160–74.
16. Pereira L, Freitas F, Fernandes V, Pereira JB, Costa MD, Costa S, Máximo V, Macaulay V, Rocha R, Samuels DC. The diversity present in 5140 human mitochondrial genomes. Am J Hum Genet. 2009;84(5):628–40. doi:10.1016/j.ajhg.2009.04.013.
17. Rosenberg NA. The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly and polyphyly in a coalescent model. Evolution. 2003;57(7):1465–77.
18. Jukes T, Cantor C. Evolution of protein molecules In: Munro HN, editor. Mammaliam Protein Metabolism. New York: Academic Press; 1969. p. 21–132.
19. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 1981;17:368–76.
20. Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol. 1993;10:512–26.
21. Kimura M. Estimation of evolutionary distances between homologous nucleotide sequences. Proc Natl Acad Sci. 1981;78(1):454–8.
22. Posada D. Using MODELTEST and PAUP* to select a model of nucleotide substitution. Curr Protoc Bioinforma. 2003;6–5.
23. Green PJ. Reversible jump markov chain monte carlo computation and bayesian model determination. Biometrika. 1995;82:711–32.
24. Pagel M, Meade A. Bayesian analysis of correlated evolution of discrete characters by reversible-jump markov chain monte carlo. Am Nat. 2006;167(6):808–25.
25. Dawid AP. The well-calibrated bayesian. J Am Stat Assoc. 1982;77(379): 605–10.
26. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. PLoS Biol. 2006;4(5):88. doi:10.1371/journal.pbio.0040088.
27. Hayasaka K, Gojobori T, Horai S. Molecular phylogeny and evolution of primate mitochondrial dna. Mol Biol Evol. 1988;5(6):626–44.
28. Gray RR, Parker J, Lemey P, Salemi M, Katzourakis A, Pybus OG. The mode and tempo of hepatitis c virus evolution within and among hosts. BMC Evol Biol. 2011;11(1):131.
29. Heled J, Drummond AJ. Bayesian inference of species trees from multilocus data. Mol Biol Evol. 2010;27(3):570–80. doi:10.1093/molbev/msp274.
30. Ogilvie HA, Heled J, Xie D, Drummond AJ. Computational performance and statistical accuracy of *beast and comparisons with other methods. Syst Biol. 2016;65(3):381–96. doi:10.1093/sysbio/syv118.
31. Gavryushkina A, Welch D, Stadler T, Drummond AJ. Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. PLoS Comput Biol. 2014;10(12):1003919. doi:10.1371/journal.pcbi.1003919.
32. Heled J, Drummond AJ. Bayesian inference of population size history from multiple loci. BMC Evol Biol. 2008;8:289. doi:10.1186/1471-2148-8-289.
33. Stadler T, Kühnert D, Bonhoeffer S, Drummond AJ. Birth-death skyline plot reveals temporal changes of epidemic spread in hiv and hepatitis c virus (hcv). Proc Natl Acad Sci USA. 2013;110(1):228–33. doi:10.1073/pnas.1207965110.