bioRxiv preprint doi: https://doi.org/10.1101/020792; this version posted June 11, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# bModelTest: Bayesian site model selection for nucleotide data

Remco Bouckaert[1,2]
remco@cs.auckland.ac.nz
1: University of Auckland, 2: Max Planck Institute

June 11, 2015

**Abstract**

bModelTest allows for a Bayesian approach to inferring a site model for phylogenetic analysis. It is based on trans dimensional MCMC proposals that allow switching between substitution models, whether gamma rate heterogeneity is used and whether a proportion of the sites is invariant. The model can be used with the set of reversible models on nucleotides, but we also introduce other sets of substitution models, and show how to use these sets of models. With the method, the site model can be inferred during the MCMC analysis and does not need to be pre-determined, as is now often the case in practice, by likelihood based methods.

## 1 Introduction

One of the choices that need 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 [16], JModelTest [15], or JModelTest2 [4] to determine the substitution model and whether to use gamma rate heterogeneity [21] and/or invariant sites. These three components form what we call the *site model*. The site model recommended by such likelihood based method is then used in the Bayesian analysis. A more elegant method is to co-estimate site model and phylogeny in a single analysis, thus preventing possible bias caused by the phylogeny assumed by the above methods. Obviously, in a Bayesian setting it is more natural to use Bayesian methods of model selection.

One way to select substitution models of DNA evolution is to use reversible jump between all possible reversible models [8], or just a nested set of model [2]. An alternative is to use stochastic Bayesian variable selection [5], though this does not address whether to use gamma rate heterogeneity or invariant sites. Wu et al. [20] 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 invariant sites.

bioRxiv preprint doi: https://doi.org/10.1101/020792; this version posted June 11, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We introduce a method that estimates all three components of the site model simultaneously, which combines selection of substitution model as with selection for rate heterogeneity model and invariant sites, all of these are using reversible jump. The method is implemented in the bModelTest package of BEAST 2 [3] with GUI support for BEAUti making it trivial 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`.

## 2 Method

Symmetric nucleotide models can be represented by a 4x4 rate matrix

$$Q = \begin{pmatrix} - & r_{ac} & r_{ag} & r_{at} \\ r_{ac} & - & r_{cg} & r_{ct} \\ r_{ag} & r_{cg} & - & r_{gt} \\ r_{at} & r_{ct} & r_{gt} & - \end{pmatrix}$$

with six rates $r_{ac}$, $r_{ag}$, $r_{at}$, $r_{cg}$, $r_{ct}$ and $r_{gt}$. A particular grouping of rates can conveniently be represented by a six figure number where each of the six numbers corresponds to one of the six rates in the order listed before respectively. Rates that are grouped have the same number. For examples, model 123456 corresponds to a model where all rates are independent. This is also known as the GTR model. Model 1211211 corresponds to the HKY model. By convention, the lowest number representing a model is used, so even though 646646 and 212212 represent the same model, we only use 121121.

There are 203 reversible models in total [8]. However, it is well known that transitions (A↔C, and G↔T mutations) are much more likely than tranversions (the other mutations) [13, 17]. Hence grouping transition rates with transversion rates is not appropriate and these rates should be treated differently. If we restrict the set of substitution model such that we only allow grouping within transitions and within transversions (with the exception of model 111111, where all rates are grouped), only 31 models are left (see Figure 1 and details in Appendix). If one is interested in using named models, we can use Jukes Cantor [9, 6] (111111), HKY [10] (121121), TN93 [18] (121131), K81 [11] (123321), TIM [14] (123341), TVM [14] (123421),and GTR [19] (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 nicely fit between TIM and GTR, leaving us with a set of 7 models (Figure 1).

The state space consist of the following parameters:

- the model number $M$,
- a variable size rate parameter (depending on model number) $R$,
- a flag to indicate whether 1 or 4 gamma categories should be used,
- a shape parameter $\alpha$, used for gamma rate heterogeneity when there are 4 gamma categories,

bioRxiv preprint doi: https://doi.org/10.1101/020792; this version posted June 11, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

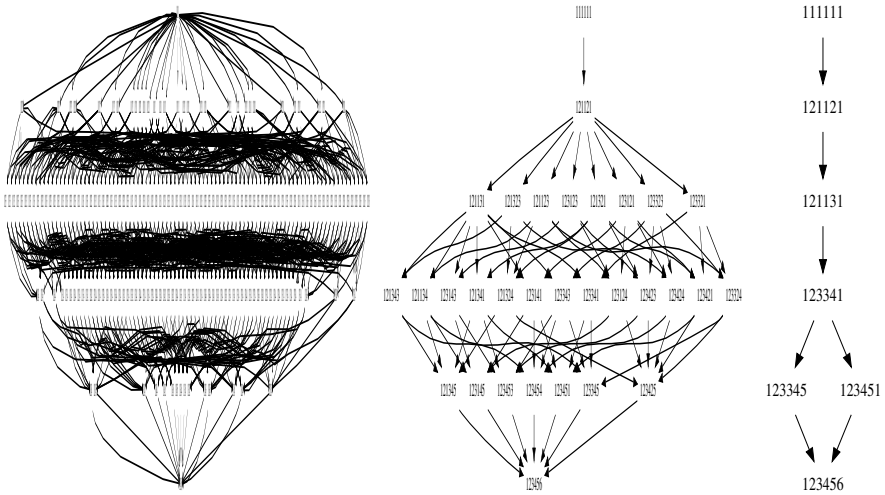


Figure 1: Model sets supported by bModelTest, all reversible models (left), transition/tranversion split models (middle), and named models (right). Arrows indicate which models can be reached by splitting a model. Note all models with the same number of groupings are at the same height.

- a flag to indicate a category for invariant sites should be used,
- the proportion of invariant sites $p_{inv}$,

Rates rates $r_{ac}$, $r_{ag}$, $r_{at}$, $r_{cg}$, $r_{ct}$ and $r_{gt}$ are determined from $M$ and $R$. Further, we restrict $R$ such that the sum of the six rates $\sum r_{..}$ equals 6 in order to ensure identifiability.

## 2.1 Prior

By default, bModelTest uses the flat Dirichlet prior on rates from [8] taking the linear transformation in account. From emperical studies [13, 17], we know that transition rates tend to be higher than transition 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) and transition rates (default exponential with mean 1).

An obvious choice for the prior on models is to use a uniform prior on all 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 groups. That way, Jukes Cantor and GTR get a prior probability of 1/6, since these are the only models with 1 and 6 groups respectively. Depending on the model set, a much lower probability is assigned to each of the other models such that the sum over models with $k$ groups is 1/6.

bioRxiv preprint doi: https://doi.org/10.1101/020792; this version posted June 11, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

For $p_{inv}$ a uniform prior with interval $(0, 1)$ is used and for $\alpha$ an exponential with mean 1. These priors only affect the posterior when the respective flags are set.

## 2.2 MCMC proposals

The probability of acceptance of a (possibly trans-dimensional) proposal [7] 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 propsoed state with respect to that of the current state.

### 2.2.1 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 differs from Huelsenbeck's approach [8] where first randomly a group is selected, then randomly a subgrouping created, since the graph based method is easier to generalise to other model sets (e.g. the one used in [12]). 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 currents 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 $\frac{n_i n_j}{n_i + n_j}$.

bioRxiv preprint doi: https://doi.org/10.1101/020792; this version posted June 11, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

#### 2.2.2 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 $\delta$ is a parameter of the proposal (0.15 appears a good number). 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.

#### 2.2.3 Birth/death proposal

Birth and death proposals set or unset the category count flag and sample a new value for $\alpha$ 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. Likewise for setting the proportion invariant flag and sampling $p_{inv}$ from the prior. The Jacobian is 1 for all these proposals.

#### 2.2.4 Scale proposal

For the $\alpha$, we use the standard scale operator in BEAST 2 [3], adapted so it only samples if the category count flag is set for $\alpha$. Likewise, for $pInv$ this scale operator is used, but only if the proportion invariant flag is set.

## 3 Results

Since implementation of the split/merge and rate exchange proposals is not straightforward, nor getting the math right for proposal ratio and Jacobian, unit tests were written to guarantee their correctness and lack of bias in proposals.

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 [1] 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.

bioRxiv preprint doi: https://doi.org/10.1101/020792; this version posted June 11, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

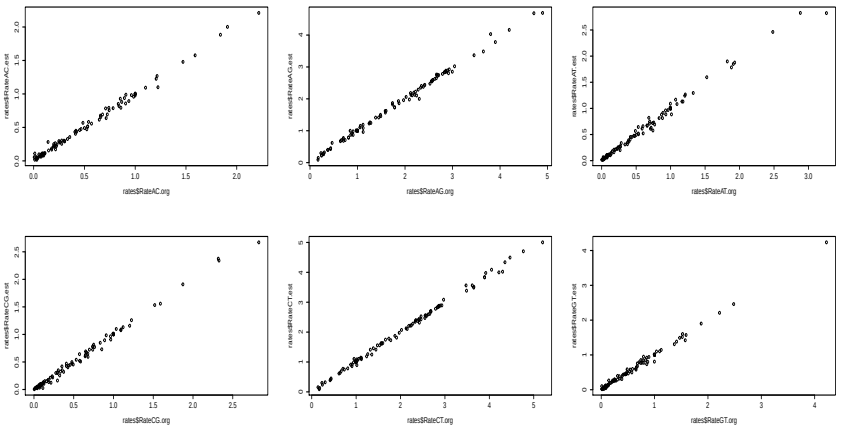


Figure 2: Rate original (horizontal) against estimates (vertical) in simulated data. In reading order, rate AC, AG, AT, CG, CT and GT.

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 low, the estimate tended to be biased upward. 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 the invariant 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 invariant. The actual proportion invariant was determined empirically by counting the number of invariant sites in the generated alignments.[2] It appears that if there is less than 60% invariant sites, adding a category to model them does not give a better fit. When invariant rates were used they had a high correlation ($R^2 = 0.77$) with that of the empirical proportion invariant.

To compare results with jModelTest 2 [4], we used gopher data (available from BEAST 2 as file examples/nexus/gopher.nex) and GTR+G is given the highest likelihood score. So, it may be tempting to just use that model. How-

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

[2] Estimated proportion invariant only take values of the proportion invariant parameter in account in the posterior sample where the invariant category was present.

bioRxiv preprint doi: https://doi.org/10.1101/020792; this version posted June 11, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

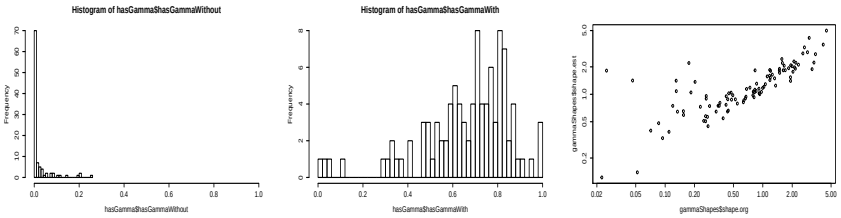


Figure 3: Proportion of time gamma rate heterogeneity is used without (left) and with (middle) rate heterogeneity used to generate the data. Right, gamma shape original (horizontal) against estimates (vertical) when rate heterogeneity is used to generate the data.

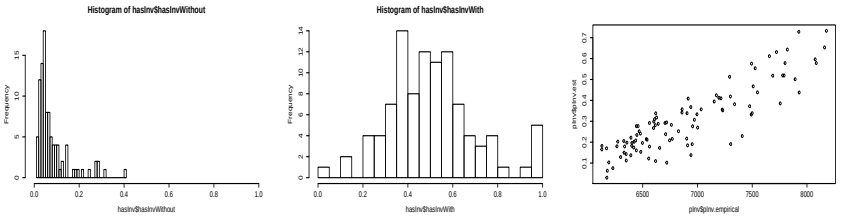


Figure 4: Proportion of time proportion invariant is used without (left) and with (middle) rate heterogeneity used to generate the data. Right, empirical proportion invariant in alignment (horizontal) against estimates (vertical) when rate heterogeneity is used to generate the data.

bioRxiv preprint doi: https://doi.org/10.1101/020792; this version posted June 11, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

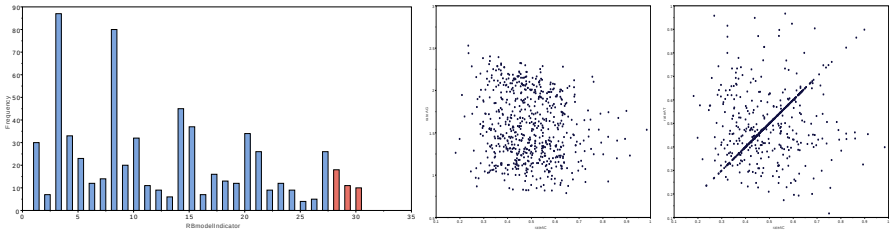


Figure 5: Model distribution for gopher data using the transition/tranversion split models (left). Numbers on x-axis correspond to models in Appendix. Rate AC vs AG (middle) and AC vs AT (right) show no correlation for the former and high correlation for the latter.

ever, with bModelTest we get a more nuanced picture (Figure 5) showing GTR only makes a small fraction of the posterior. Further, 78% of the time gamma rate heterogeneity is in use, while 60% of the posterior has an invariant site category.

The calculation of the tree likelihood consumes the bulk (>> 90%) of computational time. Note that for a category with invariant 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 [5] so that only the tree likelihood that needs updating is performing calculations.

# 4 Conclusion

bModelTest is a BEAST package which can be used in any analysis where trees are estimated based on DNA sequences, such as multi-species coalescent analysis (*BEAST), various forms of phylogeographical analyses, sampled ancestor analysis, demographic reconstruction using coalescent, birth death skyline analysis, etc. 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 determining aspects of the site model.

bioRxiv preprint doi: https://doi.org/10.1101/020792; this version posted June 11, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

## Acknowledgements

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.

## References

[1] Remco R. Bouckaert. Beastshell – scripting for bayesian hierarchical clustering. *Submitted*, 2015.

[2] Remco R. Bouckaert, Mónica Alvarado-Mora, and João Rebello Pinho. Evolutionary rates and hbv: issues of rate estimation with bayesian molecular methods. *Antiviral therapy*, 2013.

[3] Remco R. Bouckaert, Joseph Heled, Denise Kühnert, Tim Vaughan, Chieh-Hsi Wu, Dong Xie, Marc A Suchard, Andrew Rambaut, and Alexei J Drummond. Beast 2: a software platform for bayesian evolutionary analysis. *PLoS Comput Biol*, 10(4):e1003537, Apr 2014.

[4] Diego Darriba, Guillermo L Taboada, Ramón Doallo, and David Posada. jmodeltest 2: more models, new heuristics and parallel computing. *Nature methods*, 9(8):772–772, 2012.

[5] Alexei J. Drummond and Remco R. Bouckaert. *Bayesian evolutionary analysis with BEAST*. Cambridge University Press, 2015.

[6] J. Felsenstein. Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981.

[7] P.J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82:711–732, 1995.

[8] John P. Huelsenbeck, Bret Larget, and Michael E. Alfaro. Bayesian phylogenetic model selection using reversible jump markov chain monte carlo. *Molecular Biology and Evolution*, 21(6):1123–1133, June 2004.

[9] T. Jukes and C. Cantor. Evolution of protein molecules. In H.N. Munro, editor, *Mammaliam Protein Metabolism*, pages 21–132. Academic Press, New York, 1969.

[10] M. Kimura. A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16:111–120, 1980.

[11] Motoo Kimura. Estimation of evolutionary distances between homologous nucleotide sequences. *Proceedings of the National Academy of Sciences*, 78(1):454–458, 1981.

bioRxiv preprint doi: https://doi.org/10.1101/020792; this version posted June 11, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[12] Mark Pagel and Andrew Meade. Bayesian analysis of correlated evolution of discrete characters by reversible-jump markov chain monte carlo. *The American Naturalist*, 167(6):808–825, 2006.

[13] Luísa Pereira, Fernando Freitas, Verónica Fernandes, Joana B Pereira, Marta D Costa, Stephanie Costa, Valdemar Máximo, Vincent Macaulay, Ricardo Rocha, and David C Samuels. The diversity present in 5140 human mitochondrial genomes. *Am J Hum Genet*, 84(5):628–40, May 2009.

[14] David Posada. Using modeltest and paup* to select a model of nucleotide substitution. *Current protocols in bioinformatics*, pages 6–5, 2003.

[15] David Posada. jmodeltest: phylogenetic model averaging. *Molecular biology and evolution*, 25(7):1253–1256, 2008.

[16] David Posada and Keith A. Crandall. Modeltest: testing the model of dna substitution. *Bioinformatics*, 14(9):817–818, 1998.

[17] N.A. Rosenberg. The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly and polyphyly in a coalescent model. *Evolution*, 2003.

[18] K. Tamura and M. Nei. 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] Simon Tavaré. Some probabilistic and statistical problems in the analysis of dna sequences. *Lectures on mathematics in the life sciences*, 17:57–86, 1986.

[20] Chieh-Hsi Wu, Marc A Suchard, and Alexei J Drummond. Bayesian selection of nucleotide substitution models and their site assignments. *Mol Biol Evol*, 30(3):669–88, Mar 2013.

[21] Ziheng Yang. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution*, 39(3):306–314, 1994.

bioRxiv preprint doi: https://doi.org/10.1101/020792; this version posted June 11, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# 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 |
| 1 | 1 | 2 | 1 | 1 | 2 | 1 | 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 111111 : 121121
HKY 121121 : 121131
TN93 121131 : 123341
TIM 123341 : 123345, 123451
123345 : 123456
123451 : 123456
GTR 123456 :

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

bioRxiv preprint doi: https://doi.org/10.1101/020792; this version posted June 11, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/020792; this version posted June 11, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/020792; this version posted June 11, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/020792; this version posted June 11, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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, 122322, 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

bioRxiv preprint doi: https://doi.org/10.1101/020792; this version posted June 11, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/020792; this version posted June 11, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 :