



# PhyloMAd

## v1.0 Manual

by David A. Duchêne, Sebastian Duchêne, and Simon Y.W. Ho

### Contents

1. Citation, availability, and licence
2. Support
3. Overview
4. Download and installation
5. Interface and usage
  - a. Substitution model assessment
    - i. Data selection
    - ii. File formats
    - iii. Tree file selection
    - iv. Model
    - v. Test statistics
    - vi. Output
    - vii. Other options, starting, and monitoring the analysis
    - viii. Interpreting the results
  - b. Clock model assessment
    - i. Data selection
    - ii. File formats
    - iii. Test statistics
    - iv. Output
    - v. Other options, starting, and monitoring the analysis
    - vi. Interpreting the results
6. Literature cited

## 1. Citation, Availability, and licence

The current citation for PhyloMAd is the following:

Duchêne D.A., Duchêne S., Ho Y.W.S. (*in prep*) PhyloMAd: Efficient assessment of phylogenomic model adequacy.

The software PhyloMAd is provided as per the repository <https://github.com/duchene/phylomad> with no warranty of any kind. Under no circumstance are the authors or their employers responsible for any damage resulting from the use of this software. The source and documentation for this software are distributed under the GNU public licence. See <http://www.opensource.org> for details.

## 2. Support

Report any bugs as issues in the github repository, or contact David A. Duchêne directly on [david.duchene@sydney.edu.au](mailto:david.duchene@sydney.edu.au) for troubleshooting or support.

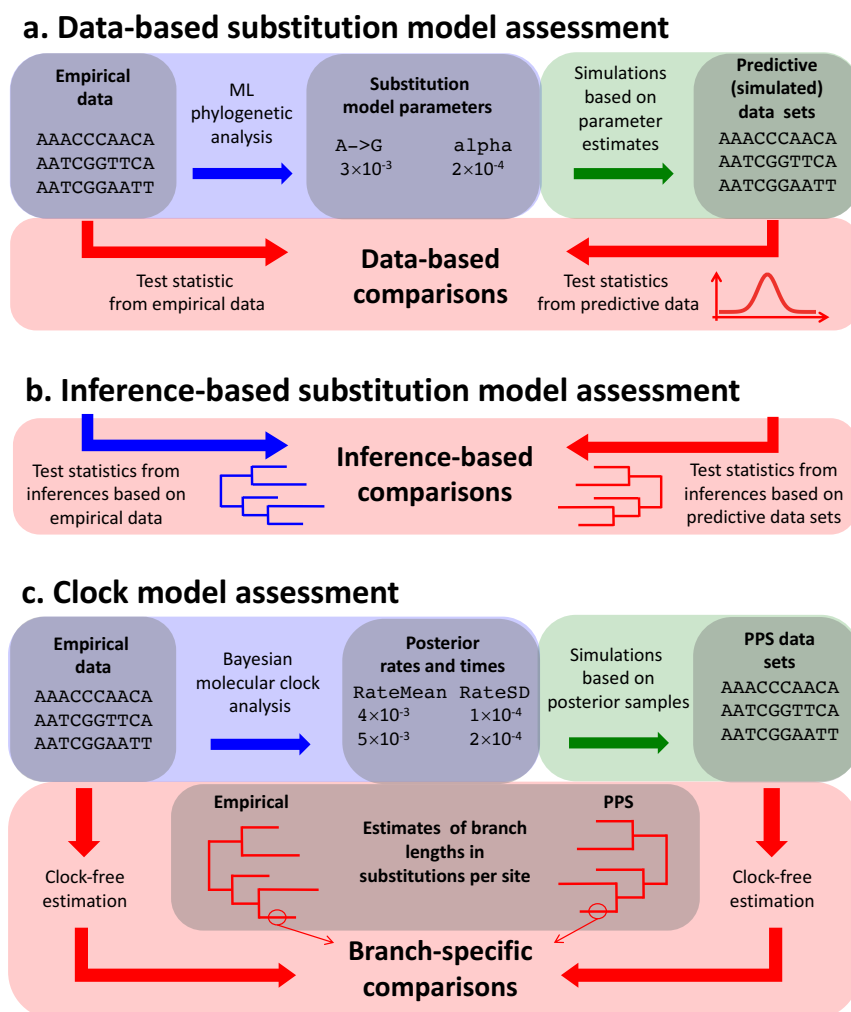
## 3. Overview

Statistical phylogenetic models are central to multiple fields of biology, but they are sometimes poor descriptions of biological reality and can produce misleading inferences. PhyloMAd assesses phylogenetic model adequacy. This is distinct from assessing model fit (e.g. using information criteria such as AIC), because it allows the assessment of the individual merits of a model, instead of comparing the relative merits across a set of candidate models. To perform this assessment PhyloMAd relies on simulating data sets that resemble the empirical data, and calculating descriptive statistics to identify differences between the data coming from the model (i.e. the simulations) and the empirical data. These simulations are also known as predictive simulations, or posterior predictive simulations when the assessment is made in a Bayesian framework. If the data originating from the model are dissimilar from the empirical data, then inferences might be misleading, and the model is rejected.

Many test statistics have been proposed for assessing model adequacy. Statistics can be classified into data-based, which describe features directly from the data (in this case assuming the data is the nucleotide or amino acid alignment; fig. 1a), and inference-based statistics (fig. 1b), which are calculated from inferences using the data and the candidate model. Multiple test statistics have been proposed for assessing substitution models, clock models (fig. 2b), and a diversity of other phylogeny-related models (e.g. the tree prior or the multi-species coalescent). Selecting and interpreting test statistics remains an active area of research, but this manual describes some of them and the interpretation of their results.

Assessment and interpretation of results in PhyloMAd is performed through a graphical user interface, and can be used for fast assessment of the adequacy of the most common substitution and clock models. The software implements a fast method of assessment using maximum likelihood, which is tailored for analyses of large multi-locus data sets and offers complete summaries and graphics of the results.

Figure 1. The four existing approaches to using posterior predictive simulations to assess model adequacy in Bayesian phylogenetics. (a) Data-based methods of assessment use characteristics of the data for model assessment, like the multinomial likelihood or the GC content. (b) Inference-based methods use characteristics of inferences for model assessment, including the tree length or the posterior probability across topologies. (c) The method for assessing clock models uses estimates from clock-free methods. The number of substitutions per site expected along each branch under the clock hierarchical model are then compared with those inferred in a clock-free analysis.



#### 4. Download and installation

PhyloMAd requires that the R statistical computing language is installed. R is freely available from the R project website.

<https://www.r-project.org/>

Download and unzip the PhyloMAd repository by clicking the following link.

<https://github.com/duchene/phylomad/archive/master.zip>

This can also be done by accessing <https://github.com/duchene/phylomad> and pressing the *Clone or download* button. Alternatively, PhyloMAd can be downloaded by opening a bash shell and typing the following. The latter option assumes that the machine has git installed.

```
git clone https://github.com/duchene/modadclocks.git
```

After download, PhyloMAd can be executed by opening the folder and double-clicking the icon according to the platform. For Mac machines press *runMac.command* and for windows press *runWin.vbs*. If needed, this will install all the required R packages, so the first time that the software is opened in a given machine might take several minutes.

If you have difficulty opening the program, you might want to try opening a bash shell, setting your directory to the PhyloMAd folder and executing the R script by hand.

```
cd path_to_PyloMAd
Rscript phylomad.Rscript
```

In mac machines, PhyloMAd will open a terminal window and log the progress. In windows, a log file with the screen progress will be saved in the main PhyloMAd folder. This file is not needed other than for checking progress, so it is safe to delete it when the program is closed.

Once you have opened PhyloMAd, start by selecting the model you wish to assess from the buttons at the left of the screen.

## 5. Interface and usage

This section will take you through the options for the model you selected to assess. These tabs will usually go from the **Data tab** to the **Other options and START** tab.

### a. Substitution model assessment

- i. *Data selection.* The **Data tab** allows the selection of the empirical data for which the model will be assessed. Alignments can be selected by pressing the Browse... button and navigating to the folder containing the data. Multiple alignments, corresponding to multiple partition subsets, can be selected simultaneously if they are in the same folder.

Figure 2. Screenshot of the Browse... button used to select data.

- ii. *File formats.* The default format is NEXUS, but Phylip and FASTA formats can be selected using the buttons below the file input.

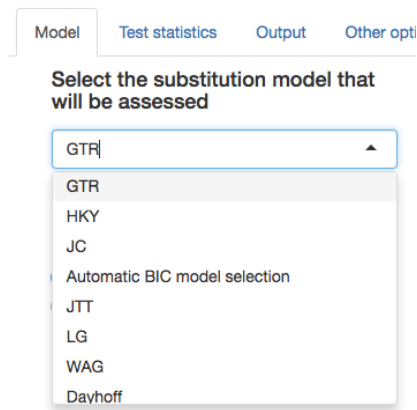
Figure 3. Screenshot of buttons to select data format.

- iii. *Tree file selection.* It might be appropriate to fix a given tree topology across the analyses. This might be the case, for example, if the topology is known from independent evidence. The tree file can be selected using the Browse button below the data selection file input. The default is to estimate the tree in every analysis, so if the tree is to be fixed it is necessary to select the tree format in this section.

Figure 4. Screenshot of optional settings to fix the tree topology.

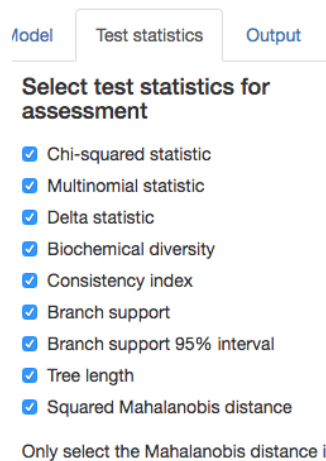
- iv. *Model.* The model to be assessed can be selected from the **Model tab** using a drop-down menu. PhyloMAd allows the assessment of nucleotide substitution models of the general time reversible family, and four common amino acid models, with the default option of including gamma-distributed rates across sites below the model drop-down menu.

Figure 5. Screenshot of dropdown menu to select the model to be assessed.



- v. *Test statistics.* The test statistics that will be calculated for model assessment can be selected from the **Test statistics tab**.

Figure 6. Screenshot of the checkboxes to select test statistics for assessment.



The following is a short description of the test statistics currently implemented in PhyloMAd:

$\chi^2$  *statistic*.—This statistic can be calculated for a table (following Sokal and Rohlf 1981) that contains the frequencies of each of the bases, for each of the taxa in the alignment. Model adequacy is assessed by computing  $\chi^2$  for data simulated under the model in question to generate a null distribution of the statistic. Importantly, this test is different from a  $\chi^2$  homogeneity test of the empirical data in that it does not use the  $\chi^2$  distribution to assess significance. Here we used the R statistical package to calculate the  $\chi^2$  test statistic of a matrix (CoreTeam 2016).

*Multinomial likelihood.*—This statistic is the likelihood of the data under a model that only describes the most general of the assumptions in the majority of substitution models: that substitution events are independent and identically distributed (i.i.d.; Goldman 1993a). A model that only describes these assumptions is not constrained by evolutionary parameters, such as probabilities of transitions and transversions, but only by the probabilities of occurrence of site patterns in the data. The likelihood of the data under this model can be calculated as the product of the frequencies of occurrence of unique sites (Reeves 1992; Goldman 1993). This statistic has been used for assessing overall model fit, primarily in a Bayesian framework (Bollback 2002; Foster 2004). A model will be rejected when using this statistic if it predicts frequencies of unique site patterns that are different from those based on the process that generated the data. For example, the JC substitution model can be rejected if the data come from more complex models of the same GTR family (Bollback 2002). This statistic can also reject models that inaccurately describe the pattern of substitution rate variation across sites. Extreme saturation of substitutions and poor alignment are other scenarios that can lead to implausible frequencies of unique sites compared with those found under common substitution models.

Importantly, sites with missing data or indels are not included when estimating the multinomial likelihood. Removing informative data is often not advisable, because sites with missing data can be informative. The test using this statistic can still be useful as a diagnostic of model adequacy in data sets with portions of missing data. For this study, we calculated the multinomial likelihood using PhyML (Guindon et al. 2010).

$\delta$  *statistic.*— This statistic is the difference between the multinomial likelihood and the maximum likelihood under a substitution model, such as those of the GTR family, and a phylogenetic tree. The  $\delta$  statistic provides insight into the “cost” of using parameters that describe the evolutionary process (Goldman 1993). A test using this statistic can, in theory, reject the substitution model if it does not adequately describe the evolutionary process. Moreover, the test can reject the substitution model if sites do not follow the i.i.d. assumption. As with the multinomial likelihood, sites with missing data or indels are not included when estimating the  $\delta$  statistic. Here we calculated the  $\delta$  statistic using PhyML.

*Biochemical diversity.*—This statistic is the mean number of distinct bases that occur at each site in the alignment (Lartillot et al.

2007). The motivation for this statistic is that it should reject the model if it does not account for the base composition across sites in the data. This test statistic is able to reject the model when there is saturation of substitutions, or if the model is unable to describe evolutionary processes across sites, such as variable rates of substitution (Lartillot et al. 2007). For this study, we used the R programming environment to calculate biochemical diversity.

*Consistency index.*— The consistency index provides a measure of homoplasy in the alignment given an estimate of the tree topology (Kluge and Farris 1969). It is calculated as the minimum possible number of substitutions in the data, as a proportion of the minimum number of substitutions required to describe the given tree topology. The motivation behind quantifying homoplasy is that it can assess whether the model can recover the amount of substitution events and phylogenetic information in the data (Lartillot et al. 2007). We calculated the consistency index using the R package phangorn (Schliep 2011). This statistic uses inferences from the data, instead of the sequences themselves as in the case of the multinomial likelihood,  $\chi^2$ , and the biochemical diversity test statistics. To implement test statistics that use inferences from the data, which include the  $\delta$  statistic, it is necessary to perform a phylogenetic analysis of the empirical data as well as of each predictive data set.

*Mean branch support.*—Some test statistics consider only the phylogenetic inferences from the data. These statistics aim specifically to assess the performance of the model in estimating parameters of interest, such as the topology and branch lengths. One such statistic is the mean node support across the inferred tree, which has been used as a test statistic for assessing substitution models in a Bayesian framework (Brown 2014). Mean branch support can be used to assess whether the amount of statistical support for a given tree is plausible under the model in question. This test statistic can detect data sets with low information content (Brown 2014). Here, we calculate the mean branch support using approximate likelihood-ratio tests in a fast maximum-likelihood approach implemented in PhyML (Anisimova and Gascuel 2006), such that inferential test statistics can be readily used on genome-scale data.

*Range in branch support.*—This statistic is based on the range of node support values across the tree (Brown 2014), and can assess whether phylogenetic information is uniformly distributed across the tree. In other words, the model can be rejected if some taxa in the



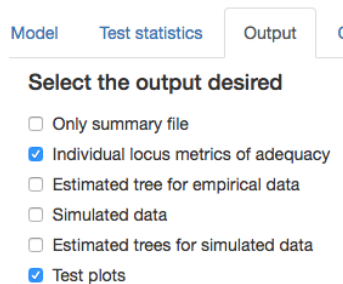
alignment have amounts of phylogenetic information that are implausible under the model in question.

*Total tree length.*—Another statistic that can be derived from phylogenetic inferences is the sum of estimated branch lengths, or total tree length (Brown 2014). This test statistic can assess whether the amount of molecular evolution inferred using the empirical data is plausible under the model. In Bayesian analyses, the tree-length statistic can identify when the branch-length prior is inappropriate, but it can fail to reject models that are mildly under-parameterized and when the data are saturated with substitutions (Brown 2014).

*Mahalanobis distance* – This statistic summarizes multiple test statistics (Mahalanobis 1936; Drummond and Suchard 2008). It is the distance between the empirical test statistics and the multivariate predictive distribution of the selected statistics. This statistic should only be selected if more than one other statistic has also been selected.

- vi. *Output.* The output can include several components. The default options will make a folder for each locus containing the metrics of adequacy for individual loci and histograms showing the results.

Figure 7. Screenshot of checkboxes to select the output desired.



Model   Test statistics   **Output**   C

**Select the output desired**

- ☐ Only summary file
- ☒ Individual locus metrics of adequacy
- ☐ Estimated tree for empirical data
- ☐ Simulated data
- ☐ Estimated trees for simulated data
- ☒ Test plots

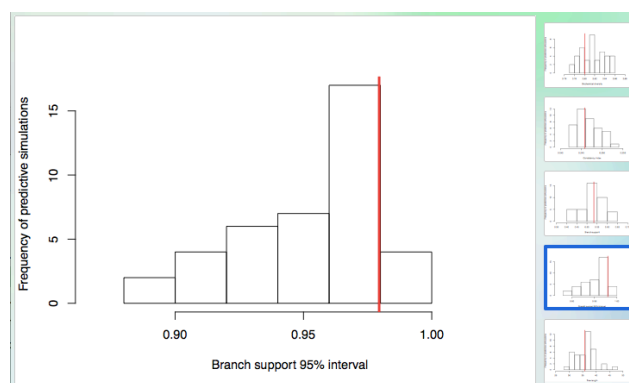
The metrics of adequacy for individual loci include two metrics of the distance between the empirical test statistic and those of the predictive data sets: one is the tail area probability (the proportion of test statistics for simulations that are lower than that calculated for the empirical data), and the other is the number of standard deviations of the predictive distribution between the mean of the distribution and the test statistic for the empirical data. The table also includes the test statistics calculated for the empirical data.

Figure 8. Example of table of metrics of adequacy for a given locus.

	chisq	multlik	delta	biochemdiv	consind	brsup	CIbrsup	trien	maha
Tail area probability	0.375	0	0.55	0.225	0	0.575	0.9	0.425	0.975
Empirical test statistic	70.2690044241849	-1059.66347	5944.14296	3.8	0.245829675153644	0.533965517241379	0.9795	35.87691978	2.74124974599074
Standard deviations from simulated distribution	-0.434195439344898	NA	0.2212590239476	-0.794394300017066	-0.25585156351048	0.133973845818382	1.01156205204035	-0.0812153816470709	-1.45204970329544

The test plots option, which is also default, leads to histograms of the test statistics calculated for predictive data, with the statistic calculated for the empirical data shown in a red line. These plots are useful for qualitative assessment of the difference between simulations and the empirical data, and is still necessary for assessment using most of the test statistics available.

Figure 9. Example of histograms for qualitative assessment of the difference between test statistics calculated for predictive data and that calculated for the empirical data set.



If the  $X^2$  statistic has been selected, PhyloMAad will also output a text file with the interpretation of the result according to a comprehensive simulation study of the statistic.

Figure 10. Example message of interpretation of results for the  $X^2$  statistic test.

```

This locus is at low risk of leading to biased inferences due to compositional heterogeneity. Homogeneous substitution
models such as those of the GTR family might provide reasonable results. It is also advisable to verify that other test
statistics do not have extreme distances from the predictive distribution.
This advice is based on the following simulations study:
Duchêne, D.A., Duchêne, S., & Ho, S.Y.W. (2017). New Statistical Criteria Detect Phylogenetic Bias Caused by
Compositional Heterogeneity. Molecular Biology and Evolution, 34(6), 1529–1534.

```

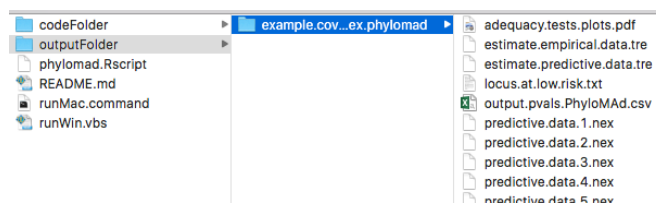
The option of producing only the summary file overrides the other output options, and leads to a single summary output file. This file contains a table in which rows contain the details of assessment of each locus. The columns include all of the table of metrics of adequacy (tail area probabilities, the test statistic calculated for the empirical data set, and the number of standard deviations from the predictive distribution). In addition, the final column contains the interpretation of the  $X^2$  statistic test in terms of the risk of bias caused by compositional heterogeneity across taxa.

Figure 11. Section of table containing a full summary of the results.

	chisq.tail.area.p	multilik.tail.area.p	delta.tail.area.p	biochemdiv.tail.area.p	consind.tail.area.p	brsup.tail.area.p	Clbrsup.tail.area.p	trien.tail.area.p	maha.tail.area.p	chisq.empirical.statistic	multilik.empirical.statistic	delta.empirical.statistic	biochemdiv.empirical.statistic	consind.emp
example.covariation.longtips.nex	0.425	0	0.675	0.375	0	0.475	0.8	0.55	1	70.2690044241849	-1059.66347	5944.14296	3.8	0.24582967

Other options will save the phylogenetic tree estimated from the empirical data, the predictive data alignments, and the phylogenetic tree estimates for each predictive data set. The following screenshot shows the complete output for an example locus (predictive alignments have been cropped).

Figure 12. Folders of PhyloMAD containing example of full output. Only a subset of the predictive data sets are shown.



- vii. *Other options, starting, and monitoring the analysis.* Final settings can be given in the **Other options and START** tab. These settings include the number of predictive data sets to be used for assessment. At least one hundred data sets should be used for assessment to be robust, but a small number can be used to test the program. It is also highly advisable to use multiple cores for analysis if allowed by the machine to make the analysis more efficient.

Figure 13. Screenshot of the boxes for setting the number of predictive simulations to be made and the number computer cores to be used for analysis.

Model
Test statistics
Output
Other options and START

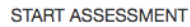
Select the number of simulations to be made

Select the number of computer cores to be used

Multi-core assessments can only be aborted at the completion of a locus assessment

Next is the button to START ASSESSMENT and follow the next prompt. After a few minutes, data should start to be stored in the output folder, and the analyses should be logged in the PhyloMAD terminal window (Mac) or in the log file in the main PhyloMAd folder (Windows).

Figure 14. Screenshot of the button that will begin the analysis.



START ASSESSMENT

- viii. *Interpreting the results.*** The aim of assessing model adequacy is to determine whether the empirical data is similar to data generated under the model, such that inferences of interest will not be misleading. The histograms that PhyloMAd provide an indication of the distance between predictive and empirical data. For quantitative metrics of the distance between the predictive data sets and the empirical data, it is perhaps most informative to report the number of standard deviations between the predictive distributions and the empirical data.

Most test statistics are not yet understood comprehensively, such that a large distance between the predictive distribution and the empirical data might not indicate a high likelihood of bias. However, PhyloMAd provides an interpretation of the  $X^2$  statistic test, and other meaningful thresholds for interpreting tests will be made available in the near future. One existing approach is to consider the model inadequate if the empirical test statistic falls outside the central 95% of the predictive distribution of statistics.

**b. Clock model assessment**

- i. *Data selection.*** In order to assess the clock model, PhyloMAd requires the output of a BEAST 2 analysis, including the log and trees files, in addition to the empirical data alignment. These can be provided by pressing the Browse... buttons in the **Data tab** and navigating the folders to select the data. PhyloMAd will extract information from these files in order to generate predictive data sets and assess the clock model. Only a single data file can be selected for each section.

In this tab, it is also possible to select a percentage of the posterior samples to be discarded as burn-in.

Figure 15. Screenshot of the Data tab when the model to be assessed is the clock (priors on branch rates and times).

- ii. *File formats.* The data alignment can be provided in one of three formats, including NEXUS, Phylip, and FASTA, using the buttons also in the **Data tab**.
- iii. *Test statistics.* The test statistics that will be calculated for model assessment can be selected from the **Test statistics tab**.

Figure 16. Screenshot of the checkboxes to select test statistics for assessment.

The following is a short description of the test statistics currently implemented in PhyloMAd for clock model assessment:

*Stemminess.*— The proportion of the inferred tree length represented by internal branches, such that stemminess is low when terminal branches are relatively long (Fiala and Sokal 1985).

*D<sub>F</sub>*.— Normalized difference between summed terminal branch lengths and total summed branch lengths in a given tree estimate.

*Tree length*.— The summed branch lengths for a given tree estimate.

*Phylogenetic imbalance*.— This is also known as “comblikeness” and is the extent to which only a single root-to-tip lineage has had diversification events. In an imbalanced tree, diversification has been imbalanced and the tree looks like a comb. The opposite is a balanced tree, and every node in the tree will contain the same number of terminal samples (tips) in each of the two descendant branches. In PhyloMAd, imbalance is calculated using the Colless index (Colless 1982).

*A index*.— To compute this index, branch lengths are estimated from the posterior predictive data alignments under a clock-free method (estimated using PhyML), such that there is a distribution of length estimates for each branch. A posterior predictive *P*-value is then calculated for each branch using the corresponding distribution obtained with the posterior predictive alignments. The *A* index for overall assessment is the proportion of branches in the phylogram from the empirical data that have lengths falling outside the 95% quantile range of those estimated from the posterior predictive data sets.

A low *A* index indicates that a large proportion of branch rates and/or times are inconsistent with the expected number of substitutions along the branches. Under ideal conditions, an *A* index of 0.95 or higher means that the clock model accurately describes the true pattern of rate variation.

- iv. *Output*. Most options in the **Output tab** behave in the same way when assessing substitutions and clock models. See section 5.b.iv and Figures 8, 9, and 11 in this manual for detailed information about the output options that overlap across types models being assessed.

One output that is unique to assessment of clock models will be produced when the *A* index is selected in the **Test statistics tab**. The *A* index produces a predictive distribution for each branch in the tree, since each branch length has been estimated from each of the predictive data sets. One of the additional files includes plots of the tree, where branches are labelled with their corresponding *P*-value and SDPD (standard deviations from the predictive distribution). Note that the trees plotted are cladograms, such that the lengths of branches have no meaning. The other additional file contains the same information but in the form of a table where the rows are the two assessment

metrics ( $P$ -value and SDPD) while columns are each of the branches. These data together with the empirical tree can be used to make further plots or analyses, and are most easily usable when loaded into R using the ape package (Popescu et al. 2012).

- v. *Other options, starting, and monitoring the analysis.* Final settings can be given in the **Other options and START** tab, and are more simple when assessing clock models compared with those of substitutions. The settings include the number of predictive data sets to be used for assessment and the number of computer cores that will be used for assessment.
- vi. *Interpreting the results.* Interpretation of results is similar to that of assessment of the substitution model. The only exception is the output of assessment using the  $A$  index, which includes results for each branch. There are two ways to interpret these branch-wise results. One is by identifying which branches fall outside the central 95% of posterior predictive branches ( $0.05 > \text{branch } P\text{-value} > 0.95$ ). The other way to interpret this is by using the branch-wise SDPD (standard deviations from the predictive distribution), and reject branches that are above a certain threshold. A threshold of 2 will be similar to that recommended for interpreting the  $P$ -value. Thresholds that are more strict (e.g. 1) or more lenient (e.g. 10) might be reasonable, but they must be reported and will only be reasonable depending on the study system.

## 6. Literature cited

- Anisimova M, Gascuel O. 2006. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst. Biol.* 55:539–552.
- Bollback JP. 2002. Bayesian model adequacy and choice in phylogenetics. *Mol. Biol. Evol.* 19:1171–1180.
- Brown JM. 2014. Detection of implausible phylogenetic inferences using posterior predictive assessment of model fit. *Syst. Biol.* 63:334–348.
- Colless DH. 1982. Review of phylogenetics: the theory and practice of phylogenetic systematics. *Syst. Zool.* 31:100–104.
- CoreTeam R. 2016. R: A language and environment for statistical computing.
- Drummond AJ, Suchard MA. 2008. Fully Bayesian tests of neutrality using genealogical summary statistics. *BMC Genet.* 9:68.
- Fiala KL, Sokal RR. 1985. Factors determining the accuracy of cladogram estimation: evaluation using computer simulation. *Evolution* 39:609.
- Foster PG. 2004. Modeling compositional heterogeneity. *Syst. Biol.* 53:485–495.
- Goldman N. 1993. Statistical tests of models of DNA substitution. *J. Mol. Evol.* 36:182–198.
- Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59:307–321.

- Kluge AG, Farris JS. 1969. Quantitative phyletics and the evolution of anurans. *Syst. Biol.* 18:1–32.
- Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol. Biol.* 7:4.
- Mahalanobis PC. 1936. On the generalised distance in statistics. *Proc. Natl. Inst. Sci. India* 12:49–55.
- Popescu A-A, Huber KT, Paradis E. 2012. ape 3.0: New tools for distance-based phylogenetics and evolutionary analysis in R. *Bioinformatics* [Internet] 28:1536–1537. Available from:  
<http://bioinformatics.oxfordjournals.org.virtual.anu.edu.au/content/28/11/1536.short>
- Reeves JH. 1992. Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. *J. Mol. Evol.* 35:17–31.
- Schliep KP. 2011. PHANGORN: Phylogenetic analysis in R. *Bioinformatics* 27:592–593.
- Sokal RR, Rohlf FJ. 1981. *Biometry*. San Francisco: W. H. Freeman