BayesRate: Bayesian estimation of diversification rates

Daniele Silvestro [1,2,3] and Jan Schnitzler [1,2]

daniele.silvestro@senckenberg.de   jan.schnitzler@senckenberg.de

version 1.4 build 20121117

Website: http://sourceforge.net/projects/bayesrate/

Please report any bugs or suggestions to the authors.

Citation: Silvestro, D., Schnitzler, J. & Zizka, G. (2011) A Bayesian framework to estimate diversification rates and their variation through time and space. BMC Evolutionary Biology, 11:311

[1] Biodiversity and Climate Research Centre (BiK-F), Senckenberganlage 25, 60325 Frankfurt am Main, Germany

[2] Department of Botany and Molecular Evolution, Senckenberg Research Institute, Senckenberganlage 25, 60325 Frankfurt am Main, Germany

[3] Diversity and Evolution of Higher Plants, Institute of Ecology, Evolution and Diversity, Goethe University, Senckenberganlage 31, 60325 Frankfurt am Main, Germany

Table of contents

License

This program is free software; you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation; either version 2 of the License, or (at your option) any later version.

This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.

System requirements and Installation

BayesRate currently runs under Mac OS X and Windows. Please note that under Windows some functions have restricted options, but this does not affect the overall capabilities of the program. BayesRate was successfully tested using R versions 2.13 - 2.15, and Python versions 2.6.4 and 2.7 (but currently does not support Python 3).

The program was written in Python based on the libraries Numpy (Ascher et al. 2001), DendroPy (Sukumaran, J. and Holder, M. T. 2010), and Scipy (Jones et al. 2001). You can download the required programs and libraries from the following websites:

Python - http://www.python.org

Numpy - http://numpy.scipy.org

DendroPy - https://github.com/jeetsukumaran/DendroPy/

A utility with graphical interface is provided within the BayesRate package (Mac: DendropyInstaller.app; Linux/Windows: DendropyInstaller.py) to facilitate the installation of Dendropy.

Scipy - http://www.scipy.org

In addition, several R packages are needed for utility functions, to process the results and generate the graphical output after the analysis. The R program and required packages (***gplots***, ***hdrcde***, ***MASS*** (Venables & Ripley 2002), and ***TeachingDemos***) are available through CRAN (http://cran.r-project.org/).

If you are working under **Windows**, please make sure that the path to **Rscript.exe** is included in the PATH environment variables (default in Mac/Linux). To do so, edit the **PATH** environment variable and add the \bin\ folder of the R installation (e.g. 'C:\Program Files\R\R-2.14.0\bin\i386'). An easy tutorial how to do that can be found for example on the Java website: http://www.java.com/en/download/help/path.xml

To start BayesRate on Windows, double-click on 'BayesRate.py'; on Linux/UNIX browse via Terminal to the BayesRate directory and type './BayesRate.py' or 'python ./BayesRate.py'. A Mac application 'BayesRate.app' is provided that can be launched by double-click.

File format

As input data, the program accepts files with distributions of ultrametric trees in NEXUS format (e.g. from a BEAST analysis). If your trees are in a different format (e.g. Newick), you will first have to convert the data into NEXUS format. Note that the NEXUS file always needs to have a TRANSLATION table and taxa names should have no spaces.

For the clade-specific analyses, an additional partition file must be provided with clade definitions and analysis settings (see also page 11 of this manual). Example files for the trees and data partitioning are provided with the program and can be found in the '*Examples*' folder of the installation directory.

Main menu

The following options are available in the main menu:

Parameter Estimation (1, 2 & 3): estimating diversification parameters under different models. These options are fast calculations of model parameters, but do not provide the model's marginal likelihood. Thus, they can not be used to compare the fit of different evolutionary models.

Marginal Likelihood (thermodynamic integration; 4, 5 & 6): calculating a model's marginal likelihood via thermodynamic integration (TDI). TDI require to run several MCMCs at different 'temperatures'. This can be carried out sequentially on a single processor, or in parallel either on a single processor (each MCMC on a different thread) or on different processors. The latter option will use up to one processor for each MCMC, if available. Note that these parallel TDI options open a new Terminal window.

Utilities (7 - 12): These options provide additional tools (e.g. Bayesian model averaging, sampling trees from a posterior distribution and plotting functions) that might be useful. The plotting functions use R to generate plots of the posterior rates and their variation through time, while calculating mean values, maximum a posteriori (MAP) estimates, and 95% credibility intervals.

Type '*help*' or '*h*' at any point while setting up an analysis to get specific information for the parameter you have to define. Use '*Ctrl-C*' to interrupt an analysis and return to BayesRate main menu. To exit the program, type '*quit*'.

NOTE: Under Windows, the 'Ctrl-C' command quits the program. When an analysis is finished type 'menu' to return to the main menu.

Output

This section describes the output provided by BayesRate. As the output depends on the analysis you run, please read this section carefully to understand where/how to find the relevant information.

BayesRate output files are by default stored in the same directory as the input file (currently there is no option to change this). The MCMC stores the sampled parameter values in a log file with the same name as the input (tree) file and the diversification model (e.g. '*file_name_PB2.log*' for a two-rate pure-birth). The only exception is the clade-specific analysis where you have to specify the name of the output file. In addition, you will find a .txt summary file that contains the MCMC settings used (e.g. '*file_name_sum_PB2.txt*').

The n-rate pure-birth model (pure-birth process with rate shifts) produces two additional log files:

1) A file that contains the marginal speciation rates for each 1 Myr bin (e.g. '*file_name_rates _PB2.log*'). Note that the speciation rate provided in the log file is the mean rate for the entire data set (across all rate shifts).

2) The posterior distribution of the temporal position of each rate shift (e.g. '*file_name_t_shift _PB2.log*').

Finally, for the thermodynamic integration (TDI), an additional folder called '*BFlogs*' is created that contains the log files of the individual chains. Log files are numbered according to the heating of the MCMC chains (usually 0 to 5 when running the default six chains). The file with the highest number (e.g. 5) contains the parameter estimates for the 'cold' chain (temperature = 1) that can be used to extract the diversification parameters.

NOTE: If you repeat an analysis with the same general settings (i.e. same diversification model), the original log files will be overwritten without warning!

If you want to compare different models of diversification (e.g. yule vs. birth-death), you have to use the 'Marginal Likelihood' option from the main menu. A text file (called e.g. '*filename_marginal.txt*') containing the model settings and the marginal likelihood will be saved in the input directory. Alternative models can then be compared using the Bayes factor test, which is defined as the ratio between their respective marginal likelihoods.

Once the log marginal likelihoods L_M are obtained via TDI, the log Bayes factor (BF) between pairs of models M_0 and M_1 can be computed as $BF = 2(M_1 - M_0)$ and interpreted following the suggestions in Kass and Raftery (1995; see table below).

2 $\log_e$ (*BF*)	Evidence against M_0
0 to 2	Not worth more than a bare mention
2 to 6	Positive
6 to 10	Strong
>10	Very strong

Acknowledgments

Thanks to Ingo Michalak for help on the Windows version of BayesRate and to Juriaan de Vos for contributing to the implementation of the key-innovation test included in this release.

References

Ascher D, Dubois PF, Hinsen K, Hugunin J, Oliphant T: Numerical Python. In UCRL-MA-128569. Livermore, CA 94566: Lawrence Livermore National Laboratory; 2001.

Jones E, Oliphant T, Peterson P: SciPy: Open source scientific tools for Python. 2001. Available from: http://www.scipy.org

Kass RE, Raftery AE: Bayes Factors. J Amer Stat Assoc 1995, 90(430):773-795.

Paradis E, Claude J, Strimmer K: APE: analyses of phylogenetics and evolution in R language. Bioinformatics 2004, 20:289-290.

Python Core Development Team: Python programming language v.2.6.4. 2010. Available from: http://www.python.org/

Rambaut A, Drummond AJ: Tracer v.1.5. 2007. Available from: http://beast.bio.ed.ac.uk/Tracer.

R Development Core Team: R: A language and environment for statistical computing v.2.13.1. R Foundation for Statistical Computing. 2011. Available from: http://www.R-project.org/

Sukumaran J and Holder MT: DendroPy: a Python library for phylogenetic computing. Bioinformatics 26, 1569-1571; 2010.

Venables WN, Ripley BD: Modern Applied Statistics with S. Fourth Edition. New York: Springer; 2002.

Command Reference

The following pages list the commands available in BayesRate together with short explanations (most of which are also available as Help in the program). Commands are written in `> Monaco regular` following a prompt (>), followed by the description. Default values (if applicable) are given in parentheses (these values will be used if no or an invalid parameter is entered).

A - Parameter Estimation

1) Speciation/Extinction rates through time

`> input tree-file (NEXUS):` Input a tree file in NEXUS format containing dated phylogenies (e.g. from BEAST output). The trees can be sub-sampled and the burnin excluded using the Utility 'Sub-sample trees'. To load a tree file simply drag-and-drop the file onto this window (by default the first tree is displayed in the terminal window).

`> please select the model of diversification:` Enter a number between 0 and 4 to select among models of diversification:

> 0: Yule process with constant rate
>
> 1: Birth-Death process with constant rates
>
> 2: Yule or Birth-Death with continuously varying rates: BOTHVAR, SPVAR, EXVAR (Rabosky & Lovette 2008)
>
> 3: n-rate pure-birth: Yule process with rate shifts (Rabosky 2006)
>
> 4: Birth-Death with incomplete taxon sampling (Yang & Rannala 1997)

`> model of diversification:`

> 0: variable speciation and extinction (default)
>
> 1: constant speciation
>
> 2: constant extinction
>
> 3: no extinction (Pure-Birth SPVAR)

`> number of rates (2) [n-rate pure-birth]:` Specify the number of different rates you want assume (minimum 2).

`> model of diversification (1) [models with taxon sampling]:` Select between Pure-Birth and Birth-Death models

`> sampling fraction (1):` Enter the proportion of species that is included in the phylogeny (e.g. 0.85; can be clade-specific)

`> number of trees (100):` Number of trees analyzed by the MCMC. Run on multiple trees to incorporate phylogenetic uncertainty in the rate estimation.

`> number per-tree MCMC iterations (100000):` Specify the number of generations the MCMC will spend sampling each tree.

`> Select option or type 'run' to start the analysis:` Select an option to customize MCMC settings or priors. Type 'run' to start the MCMC.

2) Clade-Specific Speciation/Extinction rates

> select analysis:

0) clade specific rates

1) key innovation test (Moore and Donoghue 2009)

Performs clade-specific analysis based on predefined clades (partition file is required). The option 'key innovation test' runs an estimation of the predicted species richness as described by Moore and Donoghue (2009). Please refer to the manual for more details regarding the models implemented and the format of the partition file.

`> input tree-file (NEXUS):` *see above.*

`> partition file:` To load a partition file simply drag-and-drop the file onto this window.

More information on how to setup the partition file are given in the example file and on page 11 of this manual.

3) Trait dependent diversification

This option is not yet implemented in BayesRate

B - Marginal Likelihood (thermodynamic integration)

```
> run options:
      0) sequential
      1) parallel (multi-thread)
      1) parallel (multiple processors)
```

Thermodynamic integration analyses require to run several MCMCs at different 'temperatures'. This can be carried out sequentially on a single processor (opt. 0), or in parallel either on a single processor (each MCMC on a different thread; opt. 1) or on different processors (opt. 2). The latter option will use up to one processor for each MCMC, if available.

Note that the following options will open in a new window.

`> input tree-file (NEXUS):` *see above.*

`> please select the model of diversification:` *see above.*

`> number of trees (100):` *see above.*

`> number per-tree MCMC iterations (100000):` *see above.*

`> number scaling classes (chains) (6):` Specify the number of 'heated' MCMC chains that will run in parallel to estimate the marginal likelihood.

`> Select option or type 'run' to start the analysis:` *see above.*

MCMC settings (A and B)

`> tree burnin fraction (0):` Specify the proportion or total number of trees from your input file that should be discarded as burnin.

`> tree selection (0):` Trees can be sampled: 0: consecutively, 1: evenly spaced, 2: at random. Note that the latter option should not be used for marginal likelihood estimation.

`> MCMC burnin generations (10000):` Specify the number of MCMC generations to be discarded as burnin.

`> log-to-screen frequency (10000):` Specify how frequently the state of the MCMC should be displayed on the terminal window (in generations)

`> sampling frequency (100):` Specify how often parameter values are written in the log file (e.g. every 100 generations)

`> update rates frequency:` Specify how frequently the diversification parameters are updated during the MCMC (e.g. 0.5). It might only be important if acceptance probability is very low.

`> update shift frequency:` how frequently will the shift time be updated during the MCMC. set this value to zero to fix the time of rate shift.

Set priors (rates)

`> prior on net diversification:` select the appropriate prior distribution for the net diversification rate (uniform or exponential)

`> prior on extinction fraction:` select the appropriate prior distribution for the extinction fraction (uniform or beta two-parameters)

`> lambda parameter:` Shape parameter (lambda) of the exponential distribution. The expectation is 1/lambda, but the prior allows up to infinite rates

`> alpha parameter:` Shape parameter (alpha) of the beta distribution. The expectation is alpha/(alpha+beta).

`> beta parameter:` Shape parameter (beta) of the beta distribution. The expectation is alpha/(alpha+beta).

Set constraints (shift times) Note that this option is only available for the n-rate pure-birth models

`> constraints on shift times (uniform prior):`

`lower bound of rate-shift no. ... (0):` Minimum age for the rate shift

`upper bound of rate-shift no. ... (0):` Maximum age for the rate shift

C - Utilities

7) Sub-sample trees

`> tree-file (NEXUS):` Input a tree file in NEXUS format containing dated phylogenies (e.g. from BEAST output). To load a tree file simply drag-and-drop the file onto this window.
Use 'Ctrl-C' to return to BayesRate main menu.

`> burnin fraction:` Specify the proportion or total number of trees from you input file that should be discarded as burnin.

`> number of trees:` Specify the number of trees you want to sample.

```
> Resample options:
      0) random sample
      1) evenly spaced trees
```

Trees can be sampled randomly (0) or evenly spaced (1).

8) Remove outgroups

`> tree-file (NEXUS):` Input a tree file in NEXUS format containing dated phylogenies (e.g. from BEAST output). The trees can be sub-sampled and the burnin excluded using the Utility 'Sub-sample trees'.

`> outgroups (taxa names space separated):`

NOTE: no online help is provided for this option.

9) Joint posterior - Bayesian Model Averaging

Bayesian Model Averaging (BMA) is implemented to generate a joint posterior distribution by resampling MCMC samples obtained under different models based on their relative probability.

`> log-file:` Input a log file (from 'cold' MCMC chains); type 'BMA' to stop adding log-files. To load a log file drag-and-drop the file onto this window.

`> marginal likelihood:` Enter the model marginal likelihood obtained through thermodynamic integration (use log units e.g. -124.65).

10) Plot Posterior rates

`> log-file:` Input a log file to plot posterior estimates of the diversification parameters and calculate 95% HPDs. The results will be saved in a pdf file. To load a log file simply drag-and-drop the file onto this window.

Use ‘Ctrl-C’ to return to BayesRate main menu.

11) Plot marginal rates through time (RTT)

`> log-file (_rates):` Input a log file to plot the variation of diversification rates through time. The plot will be saved in a pdf file, the mean rate and 95% HPDs per myr will be saved in a text file. To load a log file simply drag-and-drop the file onto this window.

Use ‘Ctrl-C’ to return to BayesRate main menu [MAC ONLY].

12) Plot rate shifts

`> log-file (_t_shift):` Input a log file to plot the posterior distribution of the timing of each rate-shift (if present). The plot will be saved in a pdf file, to load a log file simply drag-and-drop the file onto this window.

Use ‘Ctrl-C’ to return to BayesRate main menu [MAC ONLY].

References

Moore, BR & Donoghue, MJ: A Bayesian approach for evaluating the impact of historical events on rates of diversification. Proceedings of the National Academy of Sciences USA, 106, 4307-4312, 2009.

Example of partition file

```
# This file defines the settings for a clade-specific analysis
# Keywords (here highlighted in red) should not be modified, variables (in blue) need to be
# set by the user. All comments (in black) must be preceded by a hash ‘#’

# DEFINE PARTITIONS
name_partitions     clade_1 clade_2    # Enter the names of all clades space separated
                                       # Names must not contain spaces

output_file         example            # Specify the name of the output file
                                       # This is saved in the input file directory

# Names of the clades should match those defined above and must be followed by ':'
# To define the members of the clade enter the names of the species (space separated)

clade_1:        Taxon_01 Taxon_02 Taxon_04 Taxon_05 Taxon_06 Taxon_07 Taxon_08 Taxon_09
clade_2:        Taxon_18 Taxon_22 Taxon_20 Taxon_15 Taxon_19 Taxon_23 Taxon_01

# MODEL SETTINGS
sampling        0.5 0.7  # Enter the taxon sampling proportion for each clade
                         # When estimating predictive species richness (Moore and Donoghue 2009)
                         # the first clade works as the 'training partition' (with taxon sampling).
                         # To estimate the predictive species for the other clades (e.g. clade_2,
                         # clade_3, ...) set their respective taxon sampling to 0.

model           0 0      # boolean 0: Pure-Birth, 1: Birth-Death
                         # NOTE: for predictive species richness a PB model should be used

part_sequence   0 1      # Use same numbers to link parameters
                         # successive numbers must be used starting from 0 (e.g. 0 1 1)
                         # Note that in case of more than 2 partitions, parameters can be linked
                         # only between adjacent partitions (e.g. 0 1 1 2)

# PRIORS
prior_r         0        # shape parameter of the exponential prior on the net diversification
                         # if set to 0 a uniform prior is used

prior_a(alpha)  2        # shape parameters of the beta prior on the extinction fraction
prior_a(beta)   1        # if both set to 0 a uniform prior is used


# MCMC
trees           100      # number of trees
iterations      100000   # number per-tree MCMC iterations
sampling_freq   100      # sampling frequency
print_freq      1000     # log-to-screen frequency
burnin          10000    # MCMC burnin generations
mod_rates       1        # update rates frequency
std_updates     0.25     # standard deviation of the rate update proposals
special_rule    0        # apply special constraints (currently only available for 2 partitions)
                         # 1) link speciation rates [BD, BD] ∗ 3 parameters
                         # 2) link speciation rates [PB, BD] ∗ 2 parameters
                         # 3) link speciation rates [BD, PB] ∗ 2 parameters
                         # 4) link extinction rates [BD, BD] ∗ 3 parameters
```