babette: BEAUti 2, BEAST2 and Tracer for R

Richèl J.C. Bilderbeek[1] and Rampal S. Etienne[1]

[1]Groningen Institute for Evolutionary Life Sciences, University of Groningen, Groningen, The Netherlands

February 27, 2018

Summary

1. In the field of phylogenetics, BEAST2 is one of the most widely used software tools. It comes with the graphical user interfaces BEAUti 2, DensiTree and Tracer, to create BEAST2 configuration files and to interpret BEAST2's output files. However, when many different alignments or model setups are required, a workflow of graphical user interfaces is cumbersome.

2. Here, we present a free, libre and open-source package, `babette`: 'BEAUti 2, BEAST2 and Tracer for R', for the R programming language. `babette` creates BEAST2 input files, runs BEAST2 and parses its results, all from an R function call.

3. We describe `babette`'s usage and the novel functionality it provides compared to the original tools and we give some examples.

4. As `babette` is designed to be of high quality and extendable, we conclude by describing the further development of the package.

Keywords: computational biology, evolution, phylogenetics, BEAST2, R

1 Introduction

Phylogenies are commonly used to explore evolutionary hypotheses. Not only can phylogenies show us how species (or other evolutionary units) relate to each other, but we also estimate relevant parameters such as extinction and speciation rates. There are many phylogenetics tools available to obtain an estimate of the phylogenetic tree of a given set of species. BEAST2 [9] is one of the most widely used ones. It creates a posterior of jointly-estimated phylogenies and model parameters, from one or more DNA, RNA or amino acid alignments (see figure 1 for an overview of the workflow). It has a graphical and a command-line interface, that both need a configuration file containing alignments and model parameters. BEAST2 is bundled with BEAUti 2 [12] ('BEAUti' from now on), a desktop application to create a BEAST2 configuration file. BEAUti

has a user-friendly graphical user interface, with helpful default settings. As such, BEAUti is an attractive alternative to manual and error-prone editing of BEAST2 configuration files.

However, BEAUti cannot be called from a command-line script. This implies that when the user wants to explore the consequences of various settings, this must be done manually. This is the common workflow when using a few alignments and doing a superficial analysis of sensitivity of the reconstructed tree to model settings. For exploring many trees (for instance from simulations) and for more thorough sensitivity analysis, one would like to loop through multiple (simulated) alignments, nucleotide substitution models, clock models and tree priors. One such tool to replace BEAUti is `BEASTmasteR` [18], which focuses on morphological traits and tip-dating, but also supports DNA data. `BEASTmasteR`, however, takes hundreds of lines of R code to setup the BEAST2 model configuration and a Microsoft Excel file to specify alignment files.

BEAST2 is also associated with Tracer [22] and DensiTree [8]. Both are desktop applications to analyze the output of BEAST2, each with a user-friendly graphical user interface. Tracer's purpose is to analyze the parameter estimates generated from a BEAST 2 run. It shows, among others, the effective sample size (ESS) and time series ('the trace', hence the name) of each variable in the MCMC run. Both ESS and trace are needed to assess the strength of the inference. DensiTree visualizes the phylogenies of a BEAST2 posterior, with many options to improve the display of many phylogenies.

However, for exploring the output of many BEAST2 runs, one would like a script to collect all parameters' ESSes, parameter traces and posterior phylogenies. There is no single package that offers a complete solution, but examples of R packages that offer a partial solution are `BEASTmasteR`, `rBEAST` [23] and `RBeast` [14]. `BEASTmasteR` and `RBeast` are incomplete packages, where `rBEAST` does not aim to be a package to be used by most BEAST2 users.

Here, we present `babette`: BEAUti 2, BEAST2 and Tracer for R, which creates BEAST2 configuration files, runs BEAST2, and analyzes its results, all from an R function call. This will save time, tedious mouse clicking and reduces the chances of errors in such repetitive actions. The interface of `babette` mimics the tools it is based on. This familiarity helps both beginner and experienced BEAST2 users to make the step from those tools to `babette`. `babette` enables the creation of a single-script pipeline from sequence alignments to posterior analysis in R.

2 Description

`babette` is written in the R programming language [21] and enables the full BEAST2 workflow from an R function call, in a similar way to what BEAUti, DensiTree and Tracer do. `babette`'s main function is `run_beast2`, which configures BEAST2, runs it and parses its output. `run_beast2` needs at least the name of a FASTA file containing a DNA alignment. The default settings for the other arguments of `run_beast2` are identical to BEAUti's and BEAST2's

default settings. Per alignment, a site model, clock model and tree prior can be chosen. Multiple alignments can be used, each with its own (unlinked) site model, clock model and tree prior.

`babette` currently has 61 exported functions to set up a BEAST2 configuration file. `babette` is an alternative for a majority of BEAUti use cases. Because of BEAUti's high number of plugins, `babette` uses a software architecture that is designed to be extended. Furthermore, `babette` has 7 exported function to run and help run BEAST2. One function is used to run BEAST2, others allow the user to check if a BEAST2 configuration file is indeed valid. Finally, `babette` has 20 exported function to parse the BEAST2 output files and analyze the created posterior. `babette` gives the same ESSes and summary statistics as Tracer. The data is formatted as such, that it can easily be visualized using `ggplot2` (for a trace, similar to Tracer) or `phangorn` [24] (for the phylogenies in a posterior, similar to DensiTree).

Currently, `babette` does not replace all functionality in BEAUti, as it does not provide 3 out of 7 tree priors, nor does it support RNA alignments or use of morphological data. The many plug-ins of BEAUti are not yet supported by `babette`. `babette` does not support all command-line arguments of BEAST2, does not provide the more specialized Tracer analysis options, nor is it as feature-rich in plotting options as DensiTree. Up until now, the `babette` features implemented are those requested by users. Further extension of `babette` will be based on future user requests.

3 Usage

In R, the functions of a package need to be loaded in the global namespace first:

```
library(babette)
```

BEAUti, and likewise `babette`, needs at least a FASTA filename to produce a BEAST2 configuration file. In BEAUti, this is achieved by loading a FASTA file, then saving an output file using a common save file dialog. After this, BEAST2 needs to be applied to the created configuration file. It creates multiple files storing the posterior. These output files must be parsed by either Tracer of DensiTree. In `babette`, all this is achieved by:

```
out <- run_beast2("anthus_aco.fas")
```

This code will create a (temporary) BEAST2 configuration file, from the FASTA file with name `anthus_aco.fas` (which is supplied with the package, from [25]), using the same default settings as BEAUti, which are, among others, a Jukes-Cantor site model, a strict clock, and a Yule birth tree prior. `babette` will then execute BEAST2 using that file, and parses the output. The returned data structure, named `out`, is a list of parameter estimates (called `estimates`), posterior phylogenies (called `anthus_aco_trees`, named after the alignment's name) and MCMC operator performance (`operators`). An example of using a different site model, clock model and tree prior is:

```
out <- run_beast2(
  "anthus_aco.fas",
  site_models = create_hky_site_model(),
  clock_models = create_rln_clock_model(),
  tree_priors = create_bd_tree_prior()
)
```

This code uses an HKY site model, a relaxed log-normal clock model and a birth-death tree prior, each with their default settings in BEAUti. Table 1 shows an overview of all functions to create site models, clock models and tree priors. Note that the arguments' names `site_models`, `clock_models` and `tree_priors` are plural, as each of these can be (a list of) one or more elements. Each of these arguments must have the same number of elements, so that each alignment has its own site model, clock model and tree prior. An example of two alignments, each with its own site model, is:

```
out <- run_beast2(
  c("anthus_aco.fas", "anthus_nd2.fas"),
  site_models = list(
    create_tn93_site_model(),
    create_gtr_site_model()
  )
)
```

`babette` also uses the same default prior distributions as BEAUti for each of the site models, clock models and tree priors. For example, by default, a Yule tree prior assumes that the birth rate follows a uniform distribution, from minus infinity to plus infinity. This assumption implies that negative and positive birth rates are just as likely, where a negative birth rate is biologically impossible (note that in practice, this usually works out just fine). One may prefer an exponential distribution instead, as this would assume only positive birth rates, and makes high birth rates unlikely. To do this in `babette`:

```
out <- run_beast2(
  "anthus_aco.fas",
  tree_priors = create_yule_tree_prior(
    birth_rate_distr = create_exp_distr()
  )
)
```

Within this same example, one may specify the initial shape parameters of the exponential distribution. In BEAST2's implementation, an exponential distribution has one shape parameter: its mean, which can be set to any value with `BEAUti`. Within `babette`, to set the mean value of the exponential distribution to a fixed (non-estimated) value, do:

```
out <- run_beast2(
  "anthus_aco.fas",
  tree_priors = create_yule_tree_prior(
    birth_rate_distr = create_exp_distr(
      mean = create_mean_param(
```

```
          value = 1.0,
          estimate = FALSE
        )
      )
    )
)
```

Our initial motivation to create `babette` was that we wanted to fix the crown age of a phylogeny. BEAUti assumes that a phylogeny has a crown age that needs to be jointly-estimated with the phylogeny and other parameters. It does not allow for fixing the crown age. Without `babette`, one needs to manually edit the BEAST2 configuration file [1], which is tedious and prone to errors. Fixing the crown ages is especially useful for theoretical experiments, as this allows for one less source of variation. This is how to specify a fixed crown age with `babette`:

```
out <- run_beast2(
  "anthus_aco.fas",
  posterior_crown_age = 15
)
```

`babette` allows for the same functionality as Tracer. Tracer works on the values of the parameter estimates sampled in the BEAST2 run. This is called the "trace" (hence the name). The start of the trace is usually discarded, as an MCMC algorithm (such as used by BEAST2) first has to converge to its equilibrium. The start of the trace, called the "burn-in", will be removed, as its parameter estimates are not representative. By default, Tracer discards the first 10% of all the parameter estimates. To remove a 20% burn-in from all parameter estimates in `babette`, the following code can be used:

```
traces <- remove_burn_ins(out$estimates, burn_in_fraction =
    0.2)
```

Tracer shows the ESSes of each posterior's variables. These ESSes are important to determine the strength of the inference. As a rule of thumb, an ESS of 200 is acceptable for any parameter estimate. To calculate the effective sample sizes (of all estimated variables) in `babette`:

```
esses <- calc_esses(
  traces,
  sample_interval = 1000
)
```

Tracer displays multiple summary statistics for each estimated variable: the mean and its standard error, standard deviation, variance, median, mode, geometric mean, 95% highest posterior density interval, auto-correlation time and effective sample size. It displays these statistics per variable. In `babette`, these summary statistics are collected for all estimated parameters at once:

```
sum_stats <- calc_summary_stats(
  traces,
```

Name	Description
`run_beast2`	Run BEAST2
`create_gtr_site_model`	Create a GTR site model
`create_hky_site_model`	Create an HKY site model
`create_jc69_site_model`	Create a Jukes-Cantor site model
`create_tn93_site_model`	Create a TN93 site model
`create_rln_clock_model`	Create a relaxed log-normal clock model
`create_strict_clock_model`	Create a strict clock model
`create_bd_tree_prior`	Create a birth-death tree prior
`create_cbs_tree_prior`	Create a coalescent Bayesian skyline tree prior
`create_ccp_tree_prior`	Create a coalescent constant-population tree prior
`create_cep_tree_prior`	Create a coalescent exponential-population tree prior
`create_yule_tree_prior`	Create a Yule tree prior
`create_beta_distr`	Create a beta distribution
`create_exp_distr`	Create an exponential distribution
`create_gamma_distr`	Create a gamma distribution
`create_inv_gamma_distr`	Create an inverse gamma distribution
`create_laplace_distr`	Create a Laplace distribution
`create_log_normal_distr`	Create a log-normal distribution
`create_normal_distr`	Create a normal distribution
`create_one_div_x_distr`	Create a 1/X distribution
`create_poisson_distr`	Create a Poisson distribution
`create_uniform_distr`	Create a uniform distribution

Table 1: babette's main functions

```
  sample_interval = 1000
)
```

`babette` allows for the same functionality as DensiTree. DensiTree displays the phylogenies in a posterior at the same time scale, drawn one over one another, allowing to see the uncertainty in topology and branch lengths. Within the object `out`, the posterior phylogenies are stored as `anthus_aco_trees`, and can be plotted as such:

```
plot_densitree(out$anthus_aco_trees)
```

4 babette resources

`babette` is free, libre and open source software available from the official R package archive at `http://cran.r-project.org/src/contrib/PACKAGES.html#babette` and is licensed under the GNU General Public License v3.0. `babette` uses the Travis CI [3] continuous integration service, which is known to significantly increase the number of bugs exposed [26] and increases the speed at which new features are added [26]. `babette` has a 100% code coverage, which correlates with code quality [17, 11]. `babette` follows Hadley Wickham's style guide [27],

which improves software quality [13]. `babette` dependends on multiple packages, which are `ape` [19], `beautier` [6], `beastier` [5], `devtools` [30], `geiger` [16], `ggplot2` [28], `knitr` [32], `phangorn` [24], `rmarkdown` [4], `seqinr` [10], `stringr` [29], `testit` [31] and `tracerer` [7].

`babette`'s development takes place on GitHub [2], `https://github.com/richelbilderbeek/babette`, which accommodates collaboration [20] and improves transparency [15]. `babette`'s GitHub facilitates feature requests and has guidelines how to do so.

`babette`'s documentation is extensive. All functions are documented in the package's internal documentation. For quick use, each exported function shows a minimal example. For easy exploration, each exported function's documentation links to related functions. Additionally, `babette` has a vignette that demonstrates extensively how to use it. The GitHub documentation helps to get started, with a dozen examples of BEAUti screenshots with equivalent `babette` code.

5 Citation of babette

Scientists using `babette` in a published paper can cite this article, and/or cite the `babette` package directly. To obtain this citation from within an R script, use:

```
> citation("babette")
```

6 Acknowledgements

Thanks to Yacine Ben Chehida and Paul van Els for supplying their BEAST2 use cases. Thanks again to Paul van Els for the sharing his FASTA files for use by this package. Thanks to Leonel Herrera-Alsina, Raphael Scherrer and Giovanni Laudanno for their comments on this package and article.

References

[1] Fix starting tree. `http://www.beast2.org/fix-starting-tree`.

[2] Github. `https://github.com/`.

[3] Travis CI. `https://travis-ci.org/`.

[4] JJ Allaire, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, and Winston Chang. *rmarkdown: Dynamic Documents for R*, 2017. R package version 1.8.

[5] Richel J.C. Bilderbeek. *beastier: BEAST2 from R*. R package version 1.0, `https://github.com/richelbilderbeek/beastier`.

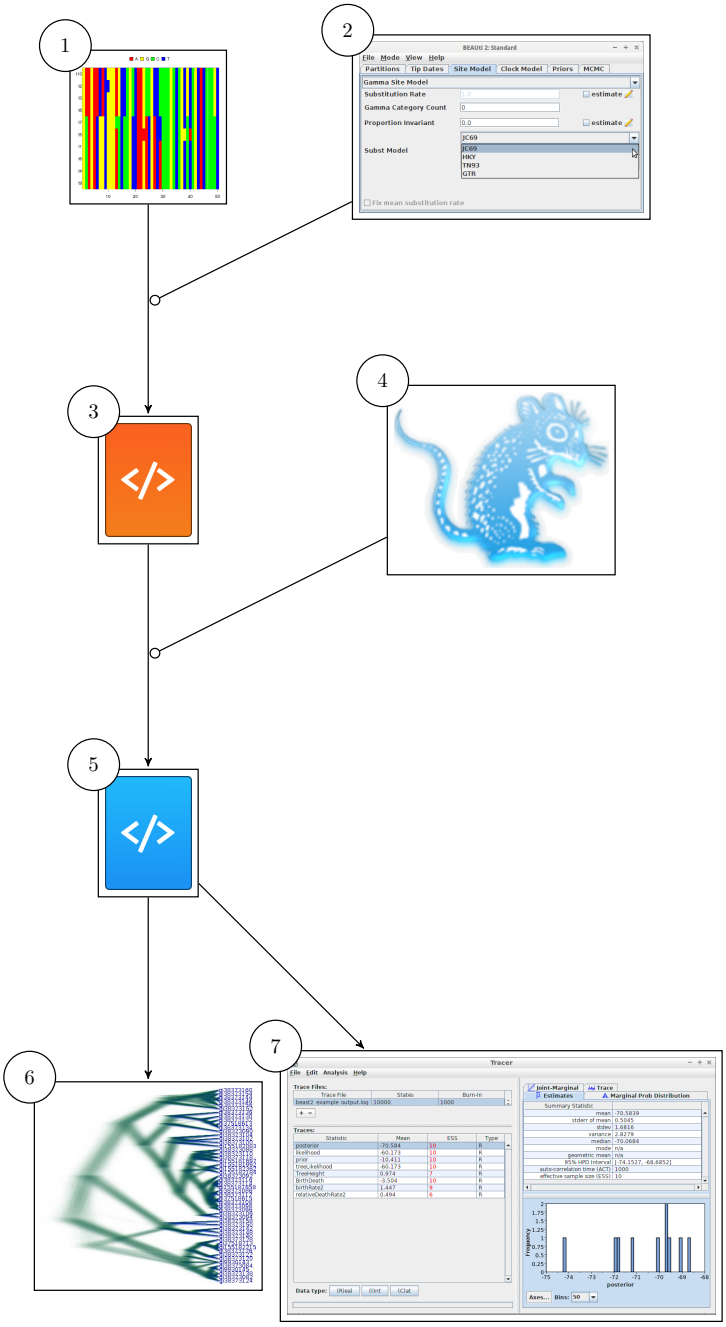

Figure 1: Workflow using GUI tools. From an alignment (1) and BEAUti (2), a BEAST2 configuration file (3) is created. BEAST2 (4) uses that file to infer a posterior, storing it in multiple files (5). These results are visualized using DensiTree (6) and Tracer (7). babette allows for the same workflow, all from an R function call.

[6] Richel J.C. Bilderbeek. *beautier: BEAUti 2 from R*. R package version 1.0, `https://github.com/richelbilderbeek/beautier`.

[7] Richel J.C. Bilderbeek. *tracerer: Tracer from R*. R package version 1.0, `https://github.com/richelbilderbeek/tracerer`.

[8] Remco Bouckaert and Joseph Heled. Densitree 2: Seeing trees through the forest. *bioRxiv*, page 012401, 2014.

[9] Remco 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, 2014.

[10] D. Charif and J.R. Lobry. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In U. Bastolla, M. Porto, H.E. Roman, and M. Vendruscolo, editors, *Structural approaches to sequence evolution: Molecules, networks, populations*, Biological and Medical Physics, Biomedical Engineering, pages 207–232. Springer Verlag, New York, 2007. ISBN : 978-3-540-35305-8.

[11] Fabio Del Frate, Praerit Garg, Aditya P Mathur, and Alberto Pasquini. On the correlation between code coverage and software reliability. In *Software Reliability Engineering, 1995. Proceedings., Sixth International Symposium on*, pages 124–132. IEEE, 1995.

[12] Alexei J Drummond, Marc A Suchard, Dong Xie, and Andrew Rambaut. Bayesian phylogenetics with beauti and the beast 1.7. *Molecular biology and evolution*, 29(8):1969–1973, 2012.

[13] Xuefen Fang. Using a coding standard to improve program quality. In *Quality Software, 2001. Proceedings. Second Asia-Pacific Conference on*, pages 73–78. IEEE, 2001.

[14] Nuno Faria and Marc A. Suchard. *RBeast*.

[15] Krzysztof J Gorgolewski and Russell Poldrack. A practical guide for improving transparency and reproducibility in neuroimaging research. *bioRxiv*, page 039354, 2016.

[16] LJ Harmon, JT Weir, CD Brock, RE Glor, and W Challenger. Geiger: investigating evolutionary radiations. *Bioinformatics*, 24:129–131, 2008.

[17] Joseph R. Horgan, Saul London, and Michael R Lyu. Achieving software quality with testing coverage measures. *Computer*, 27(9):60–69, 1994.

[18] Nicholas J. Matzke. *BEASTmasteR: R tools for automated conversion of NEXUS data to BEAST2 XML format, for fossil tip-dating and other uses*. R package version 0.2.

[19] E. Paradis, J. Claude, and K. Strimmer. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20:289–290, 2004.

[20] Yasset Perez-Riverol, Laurent Gatto, Rui Wang, Timo Sachsenberg, Julian Uszkoreit, Felipe Leprevost, Christian Fufezan, Tobias Ternent, Stephen J Eglen, Daniel SS Katz, et al. Ten simple rules for taking advantage of git and github. *bioRxiv*, page 048744, 2016.

[21] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[22] Andrew Rambaut and Alexei J Drummond. *Tracer v1.4*, 2007. Available from http://beast.bio.ed.ac.uk/Tracer.

[23] Oliver Ratmann. *rBEAST*.

[24] K.P. Schliep. phangorn: phylogenetic analysis in r. *Bioinformatics*, 27(4):592–593, 2011.

[25] Paul Van Els and Heraldo V Norambuena. A revision of species limits in neotropical pipits anthus based on multilocus genetic and vocal data. *Ibis*, 2018.

[26] Bogdan Vasilescu, Yue Yu, Huaimin Wang, Premkumar Devanbu, and Vladimir Filkov. Quality and productivity outcomes relating to continuous integration in github. In *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering*, pages 805–816. ACM, 2015.

[27] Hadley Wickham. Style guide. `http://r-pkgs.had.co.nz/style.html`. Accessed: 2017-12-20.

[28] Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009.

[29] Hadley Wickham. *stringr: Simple, Consistent Wrappers for Common String Operations*, 2017. R package version 1.2.0.

[30] Hadley Wickham and Winston Chang. *devtools: Tools to Make Developing R Packages Easier*, 2016. R package version 1.12.0.9000.

[31] Yihui Xie. *testit: A Simple Package for Testing R Packages*, 2014. R package version 0.4, `http://CRAN.R-project.org/package=testit`.

[32] Yihui Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2017. R package version 1.17.