Tutorial using BEAST v2.7.7

contraband tutorial

Rong Zhang and Fábio K. Mendes

Total-evidence dating and trait-evolution evolutionary inference using phylogenetic multivariate Brownian motion models

1 Background

Bird's-eye view. This tutorial shows how to use the `contraband` package in `BEAST 2` to model continuous trait evolution along a phylogeny with Brownian motion. Unlike methods that assume a "known", fixed tree, `contraband` lets you estimate the tempo and mode of trait evolution simultaneously with both species relationships and divergence times.

1.1 What is contraband for

In this tutorial, we will walk you through running a simple analysis with the `contraband` (**con**tinuous **tra**its **b**rowni**an** mo**d**els) `BEAST 2` package. As the name suggests, `contraband` implements Brownian motion (BM) models for the evolution of continuous traits on a phylogeny.

To understand how these models can be useful to evolutionary biologists, let's put our X-ray goggles on and look at the core of the `contraband` package: the probability density function (pdf) of the multivariate Brownian motion model – the same pdf used for a multivariate normal distribution:

$$f(\mathbf{M}|\boldsymbol{V}, \boldsymbol{y_0}) = \frac{1}{(2\pi)^{nk/2}|\boldsymbol{V}|^{1/2}} \exp\left(-\frac{1}{2}(\text{vec}(\mathbf{M}) - \boldsymbol{y_0})^{\text{T}}\boldsymbol{V}^{-1}(\text{vec}(\mathbf{M}) - \boldsymbol{y_0})\right), \quad (1)$$

This equation simply gives us the probability of observing our data $\mathbf{M}$ – that is, one or more continuous traits – given two key parameters: (i) the expected value vector (or mean vector), $\boldsymbol{y_0}$, and (ii) the variance-covariance matrix, $\boldsymbol{V}$. If you have tried a few of the other Taming the BEAST tutorials, these two parameters are the quantities whose posterior probability distributions we want to approximate via Markov Chain Monte Carlo (MCMC).

In phylogenetics, $\boldsymbol{V}$ is typically decomposed as $\boldsymbol{V} = \boldsymbol{\Sigma} \otimes \boldsymbol{T}$, where $\boldsymbol{\Sigma}$ describes the variance and covariance structure of the traits, and $\boldsymbol{T}$ represents phylogenetic relatedness. In essence, $\boldsymbol{T}$ captures the phylogeny itself – the shared evolutionary history among species.

In many software tools, especially those implemented in `R` and using frequentist methods, the phylogeny ($\boldsymbol{T}$) is not estimated but instead fixed to a tree point estimate from the literature. The downside of this approach is that the continuous trait data can only inform our estimates of trait evolution parameters, $\boldsymbol{y_0}$ and $\boldsymbol{\Sigma}$ – not the phylogeny itself.

While it is possible to take this approach in `BEAST 2` as well, its hierarchical Bayesian framework allows us to go further: we can co-estimate $\boldsymbol{T}$ (i.e., the species tree or phylogeny) together with the parameters of trait evolution. This means we can infer trait-evolution parameters **alongside** the species divergence times and phylogenetic relationships captured in $\boldsymbol{T}$. In other words, `contraband` is a tool not only for studying how continuous traits evolve, but also for estimating the topology and divergence times of phylogenies.

The estimation of divergence times using multiple types of data – for example, molecular sequences combined with discrete and/or continuous morphological traits – is known as *total-evidence dating* (TED; Ronquist et al. 2012). Among other things, `contraband` is a TED method. It is designed to help evolutionary biologists leverage continuous traits to reconstruct species evolutionary histories, including both divergence times and the tempo and mode of phenotypic evolution.

1.2 A quick peek under the hood

Later in this tutorial, you will be placing prior distributions on a series of parameters, as well as making modeling decisions related to things like the correlation between traits, for example, or the intraspecific variance in trait values. Setting up such an analysis can quickly become overwhelming, so in this section we will introduce a few implementation and statistical details to help you understand what comes next.

While it is possible to directly compute the value of equation (1) via matrix algebra, this is computationally expensive. Instead, `contraband` saves us time by using an alternative mathematical formulation (Mitov et al. 2020) and a dynamic programming algorithm. The details do not matter for this tutorial, but it is important to re-write equation (1) as:

$$f(\mathbf{M}|\boldsymbol{V}, \boldsymbol{y_0}) = f(\mathbf{M}|\Phi, \boldsymbol{y_0}, \boldsymbol{r}, \boldsymbol{\rho}, c_m, \boldsymbol{b}_m, \boldsymbol{\theta}) \tag{2}$$

You should recognize some of these terms as they have direct counterparts in models used for molecular evolution, e.g., those involved in the morphological clock model. These are [list the clock parameters here] morphological clock rate (c_m) and morphological relative branch rates ($\boldsymbol{b}_m$). This joint posterior probability density gives the posterior distribution of the time-scaled phylogenetic tree (Φ), morphological and molecular relative branch rates ($\boldsymbol{b}_m$, $\boldsymbol{b}_d$, $\boldsymbol{b}_s$) and all remaining parameters ($\boldsymbol{\theta}$) – given continuous and discrete morphology data matrices, $\mathbf{M}$ and $\mathbf{D}$, respectively, and molecular sequence alignment $\mathbf{S}$.

Other parameters, however, are unique to multivariate Brownian models, like [list those parameters here], and explain them the character values from all characters at the root of Φ ($\boldsymbol{y_0}$), a vector containing all relative character-specific evolutionary rates ($\boldsymbol{r}$), a matrix containing between-character correlation values ($\boldsymbol{\rho}$). These parameters can in principle be estimated with MCMC, but the accuracy of and uncertainty about our estimates will be a function of our data set size, which include the number of traits as well as the number of species (more details can be found in Zhang et al. 2024), as well as analysis running times.

Among the most challenging parameters to estimate are $\boldsymbol{r}$ and $\boldsymbol{\rho}$ [list parameters here] . Here, one thing that researchers can do is to ([list pre-analysis procedures, like shrinkage delta estimates]) obtain intraspecific character variation and correlation from multiple characters observed across multiple individuals within a species in the phylogeny, which amounts to ([list pre-analysis procedures, like shrinkage delta estimates]) 1) normalizing each observed character by their corresponding unbiased estimators of intraspecific variance 2) averaging an independent correlation ($\rho = 1$) and an unbiased estimate weighted by the shrinkage parameter. The assumption here is that ([list assumptions]) 1) intraspecific character variation is incremented by constant measurements from multiple individuals from a species 2) character correlations are the same across species and over time, which may be more or less justifiable depending on the data set. This is an assumption we will make in this tutorial.

Given all of the above, here is a list of the parameters we want to estimate, and for which we will need to place prior distributions on:

[add an enumerate list here with all parameters]

1) Character evolutionary rates
2) Character correlations
3) Ancestral state values
4) Tree priors
5) Clock model priors

In what follows, we will guide you through the explicit steps – including installation of dependencies and post-processing tools – that will (i) set up the analysis for inferring the above parameters, and (ii) help you process and visualize the results.

2 Programs used in this exercise

2.0.1 BEAST2 - Bayesian Evolutionary Analysis Sampling Trees2

BEAST2 (http://www.beast2.org) is a free software package for Bayesian evolutionary analysis of molecular sequences using MCMC and strictly oriented toward inference using rooted, time-measured phylogenetic trees. This tutorial is written for BEAST v2.7.7 (Bouckaert et al. 2019).

2.0.2 BEAUti2 - Bayesian Evolutionary Analysis Utility

BEAUti2 is a graphical user interface tool for generating BEAST2 XML configuration files.

Both BEAST2 and BEAUti2 are Java programs, which means that the exact same code runs on all platforms. For us it simply means that the interface will be the same on all platforms. The screenshots used in this tutorial are taken on a Mac OS X computer; however, both programs will have the same layout and functionality on both Windows and Linux. BEAUti2 is provided as a part of the BEAST2 package so you do not need to install it separately.

2.0.3 TreeAnnotator

TreeAnnotator is used to produce a summary tree from the posterior sample of trees using one of the available algorithms. It can also be used to summarise and visualise the posterior estimates of other tree parameters (e.g. node height).

TreeAnnotator is provided as a part of the BEAST2 package so you do not need to install it separately.

2.0.4 Tracer

Tracer (http://tree.bio.ed.ac.uk/software/tracer) is used to summarise the posterior estimates of the various parameters sampled by the Markov Chain. This program can be used for visual inspection and to assess convergence. It helps to quickly view median estimates and 95% highest posterior density intervals of the parameters, and calculates the effective sample sizes (ESS) of parameters. It can also be used to investigate potential parameter correlations. We will be using Tracer v1.7.2.

2.0.5 FigTree

FigTree (http://tree.bio.ed.ac.uk/software/figtree) is a program for viewing trees and producing publication-quality figures. It can interpret the node-annotations created on the summary trees by TreeAn-

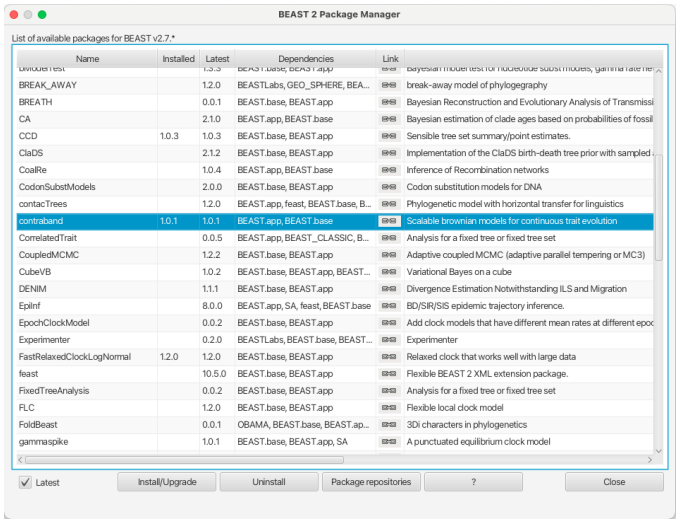

Figure 1: Download the contraband package.

notator, allowing the user to display node-based statistics (e.g. posterior probabilities). We will be using FigTree v1.4.4.

3 Practical Part I: Data preparation

4 Practical Part II: Parameter and State inference using contraband

In this tutorial we will estimate evolutionary rates, trait correlations, ancestral states and phylogenetic trees using the Brownian motion implemented in BEAST2, contraband.

The aim is to:

- Learn how to infer phylogenetic trees with continuous traits/characters
- Get to know how to choose the set-up of such an analysis
- Learn how to read the output of a “contraband" analysis

4.1 Setting up an analysis in BEAUti

4.1.1 Download contraband

First, we have to download the package contraband using the BEAUTi package manager. Go to *File >> Manage Packages* and download the package contraband.

contraband will only be available in BEAUti once you close and restart the program.

4.1.2 Loading the Carnivoran Molecular Sequences

4.1.3 Loading DATA the (Partitions)

The sequences from the *data* folder name *XXX.nexus* can be either drag and dropped into BEAUti or added using BEAUti’s menu system via *File >> Import Alignment*. Once the sequences are added, we need to specify the sampling dates.

4.1.4 Get the sampling times (Tip Dates)

Open the "Tip Dates" panel and then select the "Use tip dates" checkbox.

The sampling times are encoded in the sequence names. We can tell BEAUti to use these by clicking the *Auto-configure* button. The sampling times appear following the third vertical bar "|" in the sequence name. To extract these times, select "split on character", enter "|" (without the quotes) in the text box immediately to the right, and then select "3" from the drop-down box to the right, as shown in the figure below. Clicking "Ok" should now populate the table with the sample times extracted from the sequence names: the column **Date** should now have values between 2000 and 2002 and the column **Height** should have values from 0 to 2. The heights denote the time difference from a sequence to the most recently sampled sequence. If everything is specified correctly, the sequence with Height 0.0 should have Date 2001.9.

4.1.5 Specify the Brownian motion Model

4.1.6 Set the clock model (Clock Model)

4.1.7 Specify the priors (Priors)

Now, we need to set the priors for the various parameters of the model. You can find the parameter priors below the tree prior.

4.1.8 Specify the MCMC chain length (MCMC)

Here we can set the length of the MCMC chain and after how many iterations the parameter and trees a logged. For this dataset, 2 million iterations should be sufficient. In order to have enough samples but not create too large files, we can set the logEvery to 2000, so we have 1001 samples overall. Next, we have to save the `*.xml` file under *File >> Save as*.

4.1.9 Run the Analysis using BEAST2

Run the `*.xml` using BEAST2 or use finished runs from the *precooked-runs* folder. The analysis should take about 6 to 7 minutes.

5 Practical Part III: Post analysis

5.0.1 Analyse the log file using Tracer

First, we can open the `*.log` file in tracer to check if the MCMC has converged. The ESS value should be above 200 for almost all values and especially for the posterior estimates.

5.0.2 Make the summary tree using TreeAnnotator

Producing MCC tree

Open **TreeAnnotator** and then set the options as in the Figure 1 below. You have to specify the **Burnin percentage**, **Target tree type**, **Node heights**, **Input Tree File** and the **Output File**. Use the typed trees in the file `H3N2.H3N2.trees` as **Input Tree File**. Name output file `H3N2.mcc.tree`. After

clicking **Run** the program should summarize the trees.

5.0.3 Analyse and compare the MCC trees

5.0.4 Errors that can occur (Work in progress)

Version dated: June 23, 2025

Relevant References

Bouckaert, R, TG Vaughan, J Barido-Sottani, S Duchêne, M Fourment, A Gavryushkina, J Heled, G Jones, D Kühnert, N De Maio, et al. 2019. Beast 2.5: an advanced software platform for bayesian evolutionary analysis. *PLoS computational biology* 15: e1006650.

Mitov, V, K Bartoszek, G Asimomitis, and T Stadler. 2020. Fast likelihood calculation for multivariate Gaussian phylogenetic models with shifts. *Theor. Popul. Biol.* 131: 66–78.

Ronquist, F, S Klopfstein, L Vilhelmsen, S Schulmeister, DL Murray, and AP Rasnitsyn. 2012. A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. *Syst. Biol.* 61: 973–999.

Zhang, R, AJ Drummond, and FK Mendes. 2024. Fast bayesian inference of phylogenies from multiple continuous characters. *Systematic Biology* 73: 102–124.