DAISIEmainland: an R package for simulating an island-mainland system for macroevolution on islands

Joshua W. Lambert[1], Pedro Santos Neves[1], Rampal S. Etienne[1], Luis Valente[1,2], Richèl J.C. Bilderbeek[1]

[1] *Groningen Institute for Evolutionary Life Sciences, University of Groningen, Box 11103, 9700 CC Groningen, The Netherlands*

[2] *Naturalis Biodiversity Center, Darwinweg 2, 2333 CR Leiden, The Netherlands*

*Corresponding Author: Joshua W. Lambert, Groningen Institute for Evolutionary Life Sciences, Box 11103, 9700 CC Groningen, The Netherlands. Email: j.w.l.lambert@rug.nl.

Keywords: R, island biogeography, macroevolution, simulation

Summary

Speciation and extinction of species are two of the most fundamental processes that are investigated in the field of evolutionary biology. Ideally, one would like to study these processes in replicated isolated systems. Islands come close to this setting if colonisation is rare. However, often we cannot directly measure when an island was colonised, when a colonist species speciated, or when species went extinct. We can, however, use genetic data to reconstruct the evolutionary history of an island community. With this reconstructed data we can fit statistical models to understand macroevolutionary processes on islands. These models make simplifying assumptions in order for the fitting procedure to be tractable. One such assumption, made by, for example, the DAISIE model (from the `DAISIE` R package (Etienne et al., 2022)), is that mainland species (i.e. the pool of species that can colonise the island) cannot diversify or go extinct. `DAISIEmainland` is an R package that simulates speciation and extinction on the mainland as well as colonisation and diversification on the island. Providing a more realistic model of the island and the mainland for evolutionary biology research, `DAISIEmainland` features include: (1) simulating the evolutionary history on island species, (2) visualising that history, (3) calculating and plot summary metrics of the simulated data. The package enables the simulation of phylogenetic datasets from islands under a model more representative of biological systems to test current inference models in island biogeography.

Statement of Need

Analysis of phylogenetic data has provided many insights in evolutionary biology. Central to these advances is the R language (R Core Team, 2022) and the multitude of R packages which has facilitated the widespread utilisation of these methods (Paradis, 2006). Phylogenetic research in the domain of island biogeography was lacking until the development of the Dynamic Assembly of Island biota through Speciation, Immigration and Extinction (DAISIE) model provided several key findings for the macroevolution of island species (Valente et al., 2015, 2020). However, the performance and robustness of this island biogeography inference model is unknown when its assumptions are violated under biologically realistic scenarios. A central assumption of the DAISIE likelihood model is a static mainland pool of species that can colonise the island, i.e. simply a fixed number of mainland species that do not undergo speciation or extinction. `DAISIEmainland` relaxes this assumption by simulating phylogenetic data of island species which can be used to test

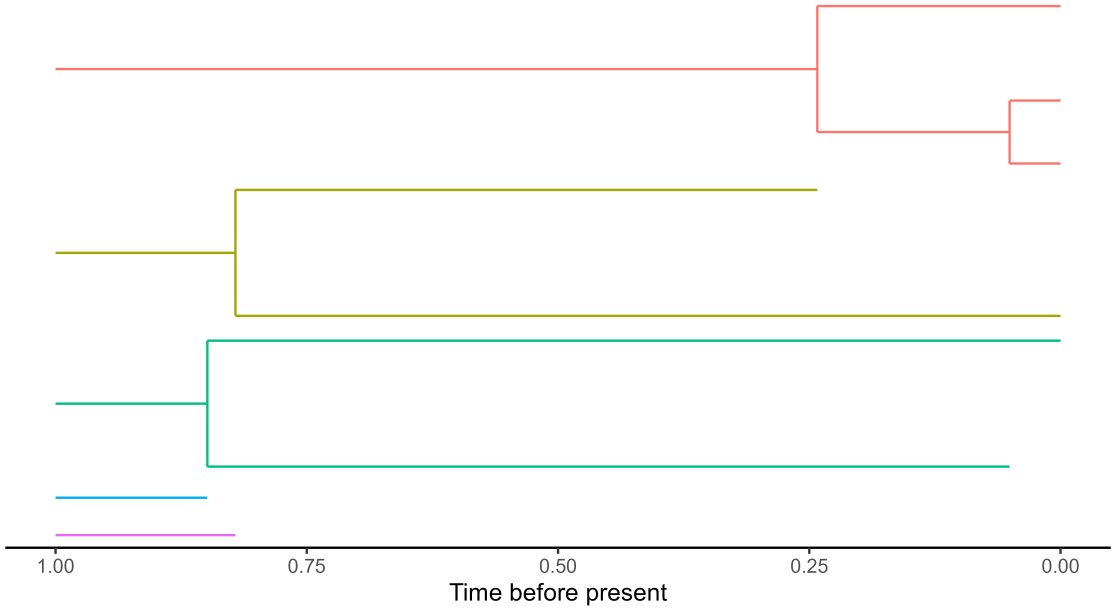

Figure 1: An example of a simulated evolutionary history of the mainland in which there are five mainland clades that form the mainland source pool of species that can colonise an island. The mainland clades are generated from a Moran process. Branching colour differentiates the mainland clades. The branching is symmetric, so when speciation occurs the original species dies and gives rise to two new species (Stadler, 2013).

whether a dynamic mainland species pool causes poor model estimation performance. The package allows for testing multiple scenarios that may be encountered by empiricists: mainland species go extinct before the present, mainland species are taxonomically known but not phylogenetically sampled, and mainland species are taxonomically undiscovered. The `DAISIEmainland` package has been applied to test the robustness of the DAISIE model (Lambert et al., 2022) (see the Inference performance chapter of the package documentation[1] for details). `DAISIEmainland` outputs data in the DAISIE format (Etienne et al., 2022), for ease of application to the `DAISIE` R package, which provides a suite of phylogenetic likelihood inference models for island biogeography.

Simulation and visualisation of the evolutionary history on islands

`DAISIEmainland` simulates events on both the mainland and the island using the Doob-Gillespie algorithm, a stochastic continuous-time process (Gillespie, 1976, 1977, 2007). The mainland is simulated under a Moran process (Moran, 1958), whereby every species extinction is immediately followed by a random species giving rise to two new species (speciation), so the number of species at any given time remains constant (Fig. 1).

The island Doob-Gillespie algorithm is run after the mainland algorithm and is altered to accommodate the dynamic mainland pool. The time-steps are bounded to not jump over changes on the mainland to ensure the present state of the system is always up-to-date regarding new species or extinct species on the mainland. The algorithm checks whether any changes have occurred on the mainland since the last time step and if so the system is updated and returned to the time at which the mainland last changed. This is valid owing to the Markov (memoryless) property of the Doob-Gillespie algorithm (Gillespie, 1976, 1977, 2007).

[1] https://joshwlambert.github.io/DAISIEmainland/inference-performance.html

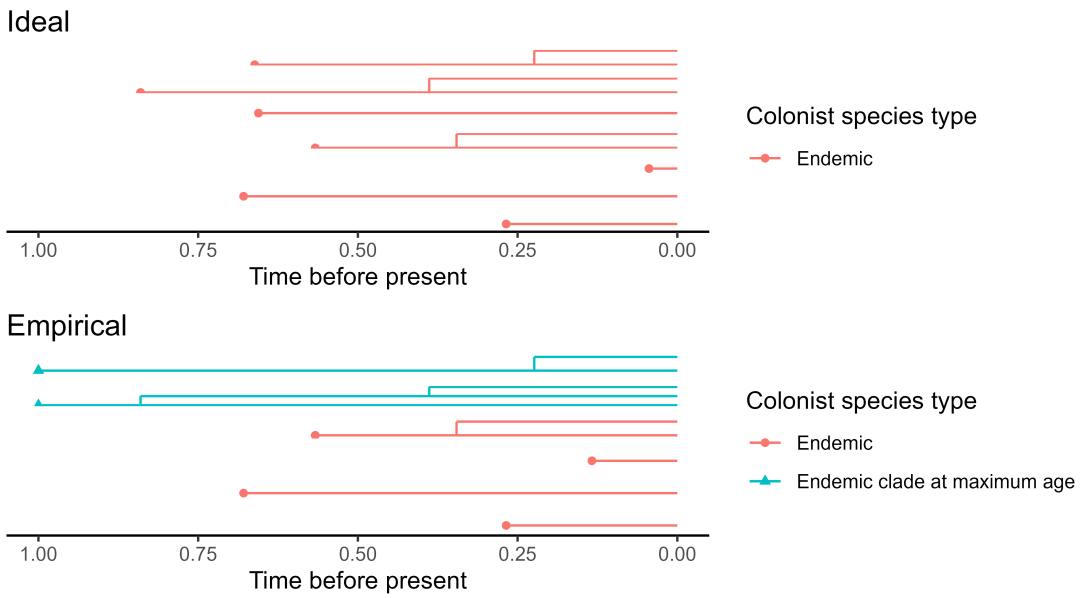

Figure 2: Simulated island colonisation and diversification data. The *ideal* data is the complete knowledge data set, where every colonisation and branching time is known exactly, without error. The *empirical* data is the incomplete knowledge data set (emulating what empirical island biogeographers would usually have access to). The branch colours represent the endemicity status of the island species, whereby endemic species only occur on the island, and non-endemic species occur on the island and mainland. The branching is asymmetric, so when speciation occurs the original species survives and gives rise to one new species (Stadler, 2013).

For both the island and mainland the timing and type of events are sampled from an exponential distribution, based on the rates of all possible events. For the mainland process, mainland extinction rate (μ_M) is the only parameter, whereas for the island there are rates of: cladogenesis (λ^c) (i.e. one island species bifurcating to form a new species on the island), island extinction (μ), colonisation (γ), and anagenesis (λ^a) (i.e. an island species becoming different from its mainland ancestor). The Doob-Gillespie samples time steps and events until the time step exceeds the total time of the simulation. The simulated data is formatted and the endemicity of each island colonist is assigned which is used in the `DAISIE` inference model (for details see the Simulation algorithm chapter of the package documentation[2]).

The `DAISIEmainland` simulation outputs two data sets: (1) contains full information of the colonisation times for all species (termed *ideal*), and (2) an incomplete information data set which resembles what an empirist would have access to (termed *empirical*). These two data sets allow for the quantification of error in estimation when the empiricist does not have access to all the data (Fig. 2). `DAISIEmainland` can plot the simulated island and mainland histories (Fig. 1 and 2; see the Simulation data visualisation chapter of the documentation[3]).

Calculate and plot summary metrics of the simulated data

The package also outputs a number of summary metrics and error metrics that are used to quantify the differences between the full information data set and the incomplete information data set simulated. These summary and error metrics include: number of species and number of colonisations on the island at the end of the simulation, the difference in normalised cumulative island colonisations through time (ΔnCTT) between *ideal* and *empirical* data, percentage of island colonisations

[2]https://joshwlambert.github.io/DAISIEmainland/simulation-algorithm.html
[3]https://joshwlambert.github.io/DAISIEmainland/simulation-data-visualisation.html

that appear to occur before the existence of the island due incomplete phylogenetic data, and percentage of species on the island that are endemic and non-endemic. There are functions for plotting each of these metrics (see Summary and error metrics visualisation chapter[4] of the package documentation).

Acknowledgements

JWL was funded through a Study Abroad Studentship by the Leverhulme Trust and an NWO VICI grant awarded to Rampal S. Etienne. PSN was funded through a FCT PhD Studentship with reference SFRH/BD/129533/2017, co-funded by the Portuguese Ministério da Ciência, Tecnologia e Ensino Superior and the European Social Fund.

References

Etienne, R. S., L. M. Valente, A. B. Phillimore, B. Haegeman, J. W. Lambert, P. S. Neves, S. Xie, R. J. C. Bilderbeek, and H. Hildenbrandt, 2022. DAISIE: Dynamical Assembly of Islands by Speciation, Immigration and Extinction. URL `https://doi.org/10.5281/zenodo.6474984`.

Gillespie, D. T., 1976. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. Journal of Computational Physics 22:403–434.

———, 1977. Exact stochastic simulation of coupled chemical reactions. The Journal of Physical Chemistry 81:2340–2361.

———, 2007. Stochastic Simulation of Chemical Kinetics. Annual Review of Physical Chemistry 58:35–55.

Lambert, J. W., P. S. Neves, R. J. C. Bilderbeek, L. Valente, and R. S. Etienne, 2022. The effect of mainland dynamics on data and parameter estimates in island biogeography. preprint, Evolutionary Biology. URL `http://biorxiv.org/lookup/doi/10.1101/2022.01.13.476210`.

Moran, P. A. P., 1958. Random processes in genetics. Mathematical Proceedings of the Cambridge Philosophical Society 54:60–71.

Paradis, E., 2006. Analysis of phylogenetics and evolution with R. Use R! Springer, New York. OCLC: ocm71298831.

R Core Team, 2022. R: A Language and Environment for Statistical Computing. URL `https://www.R-project.org/`.

Stadler, T., 2013. Recovering speciation and extinction dynamics based on phylogenies. Journal of Evolutionary Biology 26:1203–1219.

Valente, L., A. B. Phillimore, and R. S. Etienne, 2015. Equilibrium and non-equilibrium dynamics simultaneously operate in the Galápagos islands. Ecology Letters 18:844–852.

Valente, L., A. B. Phillimore, M. Melo, B. H. Warren, S. M. Clegg, K. Havenstein, R. Tiedemann, J. C. Illera, C. Thébaud, T. Aschenbach, and R. S. Etienne, 2020. A simple dynamic model explains the diversity of island birds worldwide. Nature 579:92–96.

[4] https://joshwlambert.github.io/DAISIEmainland/summary-error-metrics-visualisation.html