Simulating trajectories and phylogenies from population dynamics models with TiPS

Gonché Danesh[1,*], Emma Saulnier[1,2], Olivier Gascuel[2], Marc Choisy[3,4], Samuel Alizon[1]

[1] MIVEGEC, CNRS, IRD, Université de Montpellier
[2] Unité Bioinformatique Evolutive, C3BI/DBC USR 3756, Institut Pasteur, CNRS, Paris, France
[3] Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, UK
[4] Oxford University Clinical Research Unit, Ho Chi Minh City, Vietnam
[*] corresponding author: gonche.danesh@ird.fr

Abstract

We introduce TiPS, an R-based simulation software to generate time series and genealogies associated with a population dynamics model. The approach is flexible since it can capture any model defined with a set of ordinary differential equations (ODE), and allow parameter values to vary over time periods. Computational time is minimal thanks to the use of the Rcpp package to compile the ODEs into a program corresponding to an implementation of the Gillespie algorithm. This software is particularly suited for epidemiology and phylodynamics, where there is a need to generate numerous phylogenies for a variety of infections life cycles, and in population genetics as well.

Keywords: R package; simulation; phylogenies; compartment models; population dynamics

Introduction

Stochastic population dynamics simulations are routinely used in biology, ecology, or epidemiology [1–3]. These can be used to generate trajectories, also called time series, and genealogies that capture the relatedness between individuals. The increasing amount of genetic data is fuelling interest in linking population dynamics and genealogies because the way organisms spread can leave footprints in their genomes [4–6]. In particular, phylodynamics studies that rely on Approximate Bayesian Computing (ABC) require many simulated datasets [7–9]. Therefore, there is an increasing need for simulating trajectories as well as phylogenies rapidly and flexibly.

One of the most common methods to simulate population dynamics is Gillespie's stochastic simulation algorithm (SSA) [10], which derives from the formal result by [11] and has been implemented in a variety of programming languages. In R, packages `GillespieSSA` [12], `adaptivetau` [13], and `epimdr` [14] allow to simulate trajectories but none of these allows to simulate genealogies. Conversely, the R packages `geiger` [15], `phytools` [16], `ape` [17], and `TreeSim` can simulate phylogenies using a birth-death model but they lack population dynamics dimension and the date of the tips of the tree are always drawn at random. One exception is the R package `rcolgem` (updated to `phydynR`) that allows simulating both trajectories and phylogenies using a coalescent process. It is also available as a software package PhyDyn [18] in the BEAST2 platform [19]. Another exception is the software package MASTER [20] in BEAST2. Both are not easy to use and phydynR package's speed can be limited when simulating large phylogenies.

We introduce a flexible and easy-to-use R package to rapidly simulate population dynamics and phylogenies using a backward-in-time, i.e. coalescent, process. We performed a benchmarking analysis and found that our tool is comparable with the R package adaptivetau, but is the fastest when using our new algorithm to simulate trajectories, and is at least one order of magnitude faster than phydynR to simulate phylogenies. In Appendix, we provide additional details and the benchmarking analysis.

Methods

Structure overview

TiPS has two types of stochastic simulation outputs: population trajectories and phylogenies. These are obtained for a population-level model defined as a system of ordinary differential equations (ODE). This is performed by generating simulators in C++ and compiling them into R functions using the Rcpp package [21] for improved computational speed. The general structure of the pipeline is illustrated by the diagram in Figure 1.

Model description

We illustrate the functioning of TiPS using the SIR epidemiological compartmental model, where individuals can be susceptible (with density S), infected (I), and removed (R) [2]. The corresponding ODE system is

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\beta\ S\ I \tag{1a}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \beta\ S\ I - (\gamma + \alpha)\ I \tag{1b}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \gamma\ I \tag{1c}$$

where β is the transmission rate, γ the recovery rate, and α the virulence.

Model definition by the user

```
reactions <- c("dS = -beta*S*I",
               "dI = beta*S*I - gamma*I",
               "dR = gamma*I")
```

Building the simulator

```
# Load package
library(TiPS)

# Build and compile simulator as function
sir.simu <- build.simulator(reactions=reactions,
                            deme=c("I"))
```

Simulating trajectories

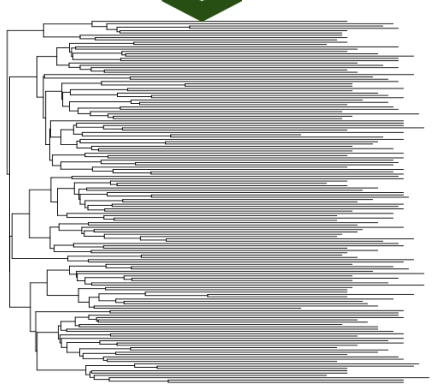
```
# Fix values of initial states, of time limits and
of parameters
x0 <- c(S=999,I=1,R=0)
times <- c(0,10)
theta <- list(gamma=0.3,beta=0.002)

# Simulate trajectory
traj <- sir.simu(paramValues=theta,
                 initialStates=x0,
                 times=time)
```

Simulating phylogenies

```
# Input : sampling dates, rooting compartment
and sampling proportion

SIR_tree <- simulate.tree(simuResults=traj,
                          dates="dates.txt",
                          root="I",
                          sampled=c("I"=1))
```



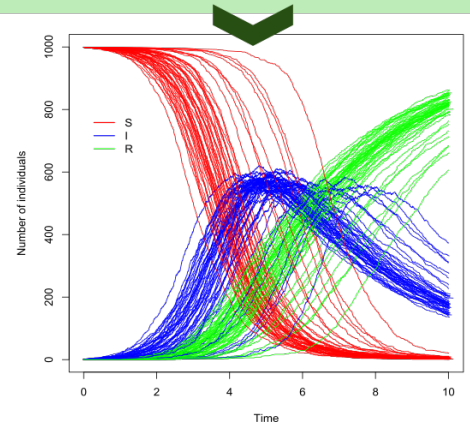


Figure 1: **Structure of the TiPS architecture.** The equations and outputs correspond to the SIR model. The functions of the R package are in blue. The simulator of trajectories built as a function is in orange. The variable *traj* in red is the output trajectory. The phylogeny is plotted using the ape R package. [17].

The model can be described as an individual-based model using a system of reactions:

$$S + I \xrightarrow{\beta IS} I + I \tag{2a}$$

$$I \xrightarrow{\gamma I} R \tag{2b}$$

$$I \xrightarrow{\alpha I} \emptyset \tag{2c}$$

where the rate of occurrence of each reaction is indicated above the transition arrow. More formally, we define three types of event reactions:

- 'Birth' reactions lead to a new individual in one of the compartments (Eq 2a),
- 'Migration' reactions correspond to the transition of an individual from one compartment to another (Eq 2b),
- 'Death' reactions correspond to the removal of an individual from the system (Eq 2c).

Simulating trajectories

TiPS uses `Rcpp` to write and compile compile into C++ a simulator that can then be used through R as a function.

Three simulation algorithms are implemented:

1. Gillespie's Direct Algorithm (GDA) [10] simulates the time until the next event by assuming that waiting times are exponentially distributed. A limitation is that its computational complexity scales linearly with the number of events and the population size.
2. Gillespie's Tau-Leap Algorithm (GTA) [22] introduces a fixed time-step during which the number of event of each type is assumed to be Poisson-distributed. This algorithm is limited when few events occur, e.g. early in an epidemics.
3. A Mixed Simulation Algorithm (MSA) that switches from GDA to GTA if over 10 iterations the time until the next event is below a threshold, and from GTA to GDA if the total number of realised event is lower than the number of possible events. For similar variations of the GDA see the next reaction method [23], the optimized direct method [24], the sorting direct method [25], or an adaptive explicit-implicit tau-leaping method [26].

The user can also input a vector of breaking time points, which allows parameter values to vary over time.

Simulating phylogenies

TiPS uses a coalescent approach [27] to simulate phylogenies based on simulated trajectories, i.e. a list of dated events (or 'reactions'), and known sampling dates (typically corresponding to observed data in ABC approaches).

First, it incorporates the sampling dates into the events of the simulated trajectory. The user needs to define which compartments can be sampled. If more than one compartment can be sampled, the user needs to specify the proportion of external nodes (i.e. leaves) associated with each class. Sampling dates are then randomly associated with a compartment label.

The tree simulation starts from the last (i.e. most recent) sampling date and progresses through the simulated trajectory backwards-in-time. Each of the four types of reactions (birth, death, migration, and sampling) can result in a modification in the simulated tree.

Multiple events may occur at the same date with GTA or the MSA and multiple sampling events may also occur at the same date (e.g. sampling campaigns). If there are multiple occurrences of the same

type of event, we assume that the number of events that lead to a change in the phylogeny can be drawn without replacement from a hypergeometric distribution. Further details can be found in the Appendix.

The output simulated phylogeny is a "phylo" class R object.

Discussion

TiPS simulates trajectories and, from these, phylogenies using a coalescent approach. The analyses in the Appendix show that it outperforms existing R packages in terms of speed when generating numerous trajectories or phylogenies. Future extensions will consist in introducing non-Markovian dynamics, spatial structure, and simulating multifurcating phylogenies.

References

[1] Otto SP, Day T. A biologist's guide to mathematical modeling in ecology and evolution. Princeton University Press; 2007.

[2] Keeling MJ, Rohani P. Modeling infectious diseases in humans and animals. Princeton University Press; 2008.

[3] Lenormand T, Roze D, Rousset F. Stochasticity in evolution. Trends in Ecology and Evolution. 2009;24(3):157–165.

[4] Grenfell BT, Pybus OG, Gog JR, Wood JLN, Janet M, Mumford JA, et al. Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. Science. 2004;303(5656):327–332.

[5] Volz EM, Koelle K, Bedford T. Viral phylodynamics. PLoS Comput Biol. 2013;9(3):e1002947.

[6] Frost SD, Pybus OG, Gog JR, Viboud C, Bonhoeffer S, Bedford T. Eight challenges in phylodynamic inference. Epidemics. 2015;10:88–92.

[7] Ratmann O, Donker G, Meijer A, Fraser C, Koelle K. Phylodynamic inference and model assessment with approximate bayesian computation: influenza as a case study. PLoS Comput Biol. 2012;8(12):e1002835.

[8] Gascuel F, Ferrière R, Aguilée R, Lambert A. How Ecology and Landscape Dynamics Shape Phylogenetic Trees. Systematic Biology. 2015;64(4):590–607.

[9] Saulnier E, Gascuel O, Alizon S. Inferring epidemiological parameters from phylogenies using regression-ABC: A comparative study. PLOS Computational Biology. 2017 Mar;13(3):e1005416.

[10] Gillespie DT. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. Journal of Computational Physics. 1976;22(4):403–434.

[11] Kurtz TG. Solutions of ordinary differential equations as limits of pure jump markov processes. Journal of Applied Probability. 1970;7(1):49–58.

[12] Pineda-Krch M. GillespieSSA: implementing the stochastic simulation algorithm in R. Journal of Statistical Software. 2008;25(12):1–18.

[13] Johnson P. adaptivetau: efficient stochastic simulations in R. R Package Version. 2014;.

[14] Bjørnstad ON. Epidemics. Models and Data using R. Springer; 2018.

[15] Pennell M, Eastman J, Slater G, Brown J, Uyeda J, Fitzjohn R, et al. geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. Bioinformatics. 2014;30:2216–2218.

[16] Revell LJ. phytools: An R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution. 2012;3:217–223.

[17] Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2019;35:526–528.

[18] Volz EM, Siveroni I. Bayesian phylodynamic inference with complex models. PLOS Computational Biology. 2018;14(11):e1006546.

[19] Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, et al. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. PLoS Computational Biology. 2014;10(4):1–6.

[20] Vaughan TG, Drummond AJ. A stochastic simulator of birth-death master equations with application to phylodynamics. Molecular Biology and Evolution. 2013;30(6):1480–1493.

[21] Eddelbuettel D, Francois R. Rcpp: Seamless R and C++ Integration. J Stat Software. 2011;40(1):1–18.

[22] Gillespie DT. Approximate accelerated stochastic simulation of chemically reacting systems. Journal of Chemical Physics. 2001;115(4):1716–1733.

[23] Gibson MA, Bruck J. Efficient exact stochastic simulation of chemical systems with many species and many channels. Journal of Physical Chemistry A. 2000;104(9):1876–1889.

[24] Cao Y, Li H, Petzold L. Efficient formulation of the stochastic simulation algorithm for chemically reacting systems. Journal of Chemical Physics. 2004;121(9):4059–4067.

[25] McCollum JM, Peterson GD, Cox CD, Simpson ML, Samatova NF. The sorting direct method for stochastic simulation of biochemical systems with varying reaction execution behavior. Computational Biology and Chemistry. 2006;30(1):39–49.

[26] Cao Y, Gillespie DT, Petzold LR. Adaptive explicit-implicit tau-leaping method with automatic tau selection. Journal of Chemical Physics. 2007;126(22):1–10.

[27] Kingman JFC. The coalescent. Stochastic processes and their applications. 1982;13(3):235–248.