Usage Notes

Tree shapes based on simulated trees using phase-type distributions

Here, we examine tree shapes via Aldous' β statistic (Aldous 1996) and the γ statistic (Pybus and Harvey 2000) based on simulated trees generated using age-dependent speciation rates and a constant exponential extinction rate. Similarly, we also conduct a simulation where trees are generated using a constant exponential speciation rate and age-dependent extinction rates. For both scenarios, we suppose that the age-dependent rates are decreased with species age using Example 1 in the main manuscript and implemented in "DecreasingPH_SpeciationExtinction_fixedZ.R". For this simulation, we generated 300 reconstructed simulated trees with 100 extant taxa under symmetric and asymmetric speciation modes using TreeSimGM package (Hagen and Stadler 2018) on *R*. For each scenario, we provide the values of the β statistic estimated from a set of trees as explained in the manuscript. If you would like to skip to the result of the simulations as presented in Figure 5 in the manuscript, you can load "DecreasingPH_Workspace_fixedZ.RData" and plot the figure. Code and workspace are available inside "tree_shapes_based_on_simulated_trees_using_phase-type_distributions.zip".

Estimate of the β statistic based on set of trees

"Sim_PH_Yule_Treesets.R" is used to generate a set of 1000 trees without extinction under Yule process and a set of 1000 trees under phase-type distribution for times to speciation event. The generated sets are stored in "Simulated_PH_Treesets.RData" and "Simulated Yule Treesets.RData". "PH_Treesets_Ind_Beta_CI.R" takes "Simulated_PH_Treesets.RData" as the input and is used to compute individual beta values from each tree in a given set of trees generated using phase-type distribution and treeset beta values from each tree set. The confidence interval for the treeset beta estimates is also included. More details can be found Section 2.3 and Section 3.1 of the main manuscript. The results are stored in "PH_Individual_Beta.RData" and "PH_Treesets_Beta_and_CI.RData", and correspond to Figure 4a,b in the manuscript. Similar procedure is used for sets of trees generated under Yule process where "Yule_Treesets_Ind_Beta_CI.R" takes "Simulated_Yule_Treesets.RData" as the input and their results are stored in "Yule_Individual_Beta.RData" and "Yule_Treesets_Beta_and_CI.RData". They correspond to Figure 4c,d in the main manuscript. These figures are generated using "Plot_CI_PH_Treeset_Ind_Beta.R" and "Plot_CI_Yule_Treeset_Ind_Beta.R". Code and data are stored at "estimate_of_the_beta_statistic_based_on_set_of_trees.zip".

Fitting parameters of phase-type distribution using simulated branch lengths

"Fit_PHdist_SimBranchLengths.R" is the main code where we simulate a set of 50 trees, each with 50 extant tips, and times to speciation are drawn from a phase-type distribution with known parameters. We consider three different simulation schemes: 1) simulate a set of 50 trees without extinction, 2) simulated a set of 50 trees with a constant exponential extinction rate of 0.1, 3) simulate a set of 50 trees with a constant exponential extinction rate of 0.4. Generated trees under scheme (1)-(3) are stored in "Sim_PH_Tree_NoExt.RData", "Sim_PH_Tree_Ext_0.1.RData", "Sim_PH_Tree_Ext_0.4.RData", respectively. For each simulation, we gather the branch lengths from their reconstructed trees to fit parameters values of a Coxian phase-type distribution shown in Example 1 in the manuscript. We use the built-in optim package on *R* with “L-BFGS-B” method (Byrd et al. 1995) to find parameter values that maximises the probability of observing a set of trees described in Equation 14 in the manuscript. The fitted parameter values under scheme (1)-(3) are stored in “Fitted_PH_Param_NoExt.RData”, “Fitted_PH_Param_Ext_0.1.RData”, “Fitted_PH_Param_Ext_0.4.RData”, respectively. Using fitted parameter values from each setup, we simulate trees with the same number of tips as in the simulated data and compare their distribution of branch lengths with that of the simulated data. These fitted trees under scheme (1)-(3) are stored in “Fitted_PH_Tree_NoExt.RData”, “Fitted_PH_Tree_Ext_0.1.RData”, “Fitted_PH_Tree_Ext_0.4.RData”. The results of this simulation correspond to Figure 6, 7, and 8 in the manuscript. We also apply the two-sample Kolmogorov-Smirnov tests to compare panel (b) and (c) in those figures. The result can be seen on Table 1 of the manuscript. Fitted and simulated trees are stored inside a folder at “fitting_parameters_of_phase-type_distribution_using_simulated_branch_lengths.zip”.

Model selection based on empirical phylogenies

Using the squamate reconstructed tree, stored in “Squamates.txt”, (Zheng and Wiens 2013) and the angiosperm reconstructed tree, stored in “Angiosperms.tre”, (Zanne et al. 2014), we use the Dendroscope 3 (Huson and Scornavacca 2012) to extract the following clades from the squamate tree: Gekkota (stored in “Gekkota.txt”), Iguania (stored in “Iguania.txt”), and Anguimorpha (stored in “Anguimorpha.txt”). “Model_Selection_Squamate_Tree.R” uses branch lengths from the squamate tree and those clades to fit parameter values of nine different distributions, namely the constant rate Birth-Death (crBD) model from Nee et al. (1994b) (stored in “Fitted_crBD” folder), the Coxian PH distribution with decreasing rate described in Example 1 (stored in “Fitted_Coxian_with_Decreasing_Rate” folder), the Coxian PH distribution with increasing rate described in Example 2 (stored in “Fitted_Coxian_with_Increasing_Rate” folder), the general Coxian distribution defined in Definition 2 with four, five, and six non-absorbing states (stored in “Fitted_General_Coxian” folder), exponential distribution (stored in “Fitted_Exponential_Distribution” folder), and Weibull distribution (stored in “Fitted_Weibull_Distribution” folder). Then, we perform model selection via AIC (Anderson and Burnham 2004) to select the best model. More details can be found in Section 4 of the main manuscript. The result corresponds to Table 2 in the manuscript. We apply the same procedure using the following clades on the angiosperm tree: Monocotyledoneae (stored in “Monocotyledoneae.txt”), Magnoliidae (stored in “Magnoliidae.txt”), Superasteridae (stored in “Superasteridae.txt”), and Superrosidae (stored in “Superrosidae.txt”). The result corresponds to Table 4 in the manuscript. We also perform absolute goodness-of-fit by comparing empirical branch length distribution from each clade from both trees with their fitted branch length densities using Kolmogorov-Smirnov test (Table 3 and Table 5), shown in Figure 9 and 11. The main code also takes the best-fitting parameter values from the general Coxian distribution with four non-absorbing states (stored in “Angiosperm_Tree” and “Squamate_Tree” folders) to plot the hazard function for each clade in both trees (shown in Fig.10 and 12). Code and data are stored at “model_selection_based_on_empirical_phylogenies.zip”.

Supplementary Codes

“gn(i)_equiv_large_n.R” are used to generate Figure 13. Code is stored at “supplementary_codes.zip”.

References

Aldous DJ (1996) Probability distributions on cladograms. In: Random discrete structures, Springer, pp 1–18

Anderson D, Burnham K (2004) Model selection and multi-model inference. Second NY: Springer-Verlag 63(2020):10

Byrd RH, Lu P, Nocedal J, Zhu C (1995) A limited memory algorithm for bound constrained optimization. SIAM J Sci Comput 16(5):1190–1208

Hagen O, Stadler T (2018) TreeSimGM: simulating phylogenetic trees under general Bellman–Harris models with lineage-specific shifts of speciation and extinction in R. Methods Ecol Evol 9(3):754–760

Huson DH, Scornavacca C (2012) Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst Biol 61(6):1061–1067

Nee S, May RM, Harvey PH (1994b) The reconstructed evolutionary process. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences 344(1309):305–311

Pybus OG, Harvey PH (2000) Testing macro–evolutionary models using incomplete molecular phylogenies. Proc Royal Soc B 267(1459):2267–2272

Zanne AE, Tank DC, Cornwell WK, Eastman JM, Smith SA, FitzJohn RG, McGlinn DJ, O’Meara BC, Moles AT, Reich PB, et al. (2014) Three keys to the radiation of angiosperms into freezing environments. Nature 506(7486):89–92

Zheng Y, Wiens JJ (2016) Combining phylogenomic and supermatrix approaches, and a time-calibrated phylogeny for squamate reptiles (lizards and snakes) based on 52 genes and 4162 species. Molecular phylogenetics and evolution 94:537–547