# INFERRING THE DRIVERS OF SPECIES DIVERSIFICATION

## USING STATISTICAL NETWORK SCIENCE

FRANCISCO RICHTER

# Inferring the drivers of species diversification

using statistical network science

Francisco Richter

Cover design: Jorge Peña

university of groningen

# Inferring the drivers of species diversification

## using statistical network science

**PhD thesis**

to obtain the degree of PhD at the
University of Groningen
on the authority of the
Rector Magnificus Prof. C. Wijmenga
and in accordance with
the decision by the College of Deans.

This thesis will be defended in public on

Friday 23 Apr 2021 at 14:30 hours

by

**Francisco Javier Richter Mendoza**

born on 08 August 1986
in Las Condes, Chile

**Promotors**
Prof. E. C. Wit
Prof. R. S. Etienne

**Assessment Committee**
Prof. Alexei Drummond
Prof. Marco Grzegorczyk
Prof. Veronica Vinciotti

To Clemente and Manu

# CONTENTS

# 1

# INTRODUCTION

*All good things are wild and free.*

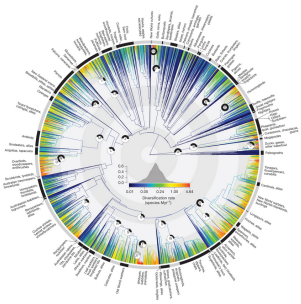
Henry David Thoreau

Species evolution is a complex process that has been studied formally for more than 200 years (Ragan, 2009) but, even today, it presents us with more questions than answers. In its essence evolutionary biology focuses on two main questions: how and why did species diversify into the enormous variety of diversity that we see today? The first is a descriptive study, whereas the second is an explanatory endeavour.

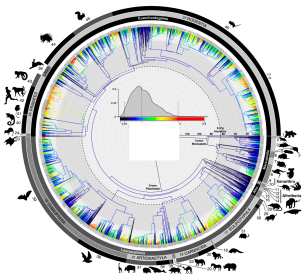
To answer how species diversified, novel DNA sequencing techniques that were developed in the early 1990s and that were perfected in the past two decades, have been vital. Their precision and high-throughput nature have made inference of large phylogenies possible. Some examples of large scale phylogenies are the tree of birds (Jetz *et al.*, 2012), the tree of mammals (Upham *et al.*, 2019), and the recently released tree of plants (Ramírez-Barahona *et al.*, 2020). Continuous efforts are made to complete the tree of life (Hug *et al.*, 2016; Hedges *et al.*, 2015a), although the accuracy of these large trees remains debated (Patel *et al.*, 2013).

Assuming, as a starting point, that current descriptions on how species diversified, i.e, the underlying phylogenies, are correct, the field of research on the underlying macroevolutionary mechanisms that drive such processes is relatively young but growing (Pagel, 1999). New advancements in phylogenetics are bringing insights into hypotheses that were difficult to test quantitatively before. The aim of this thesis is to contribute to the development of methodologies to answer the question: "What are the drivers of species diversification processes?"

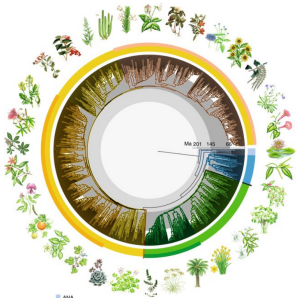
Species diversification is a highly complex process, comprising constantly emerging effects at multiple scales interacting with each other. That is a general reason why finding a general quantitative method for analysing the drivers of biodiversity has remained elusive. This thesis lies at the intersection of macroevolution and statistical modelling;



**(a)** Phylogeny of Birds provided by Jetz et. al. (Jetz *et al.*, 2012) in 2012.



**(b)** Phylogeny of Mammals provided by Upham et. al. (Upham *et al.*, 2019) in 2019.



**(c)** Phylogeny of plants provided by Ramirez et. al. (Ramírez-Barahona *et al.*, 2020) in 2020.

**Figure 1.1 |** Three large phylogenies that have recently been published.

we here present a general approach for quantitative inference over a flexible class of diversification models.

## 1.1. SPECIES DIVERSIFICATION MODELS

The theory of species diversification models (SDM) and the first mathematical theory of macroevolution date back to the 1920s when Yule started to develop models for diversification (Aldous, 2001). Yule characterised the evolutionary process as a combination of several stochastic processes, governed by speciation rates. Kendall *et al.* (1948) generalised Yule's results providing formulas for a process with constant speciation and extinction rates as well as diversification rates varying as a function of time. He also provided explicit expressions for the survival probability of the process. Much progress was achieved in the second half of the last century (Gould *et al.*, 1977; Stanley, 1973; Raup *et al.*, 1973; Reynolds, 1973, e.g.), where mathematical derivations were provided to quantify the effects of ecological dynamics on evolutionary processes described by full phylogenies, but the real application of these methods remained elusive because of the poverty of the fossil record leading to a lack of information on extinct branches. It was not until the 1990s when Nee et. al. presented their seminal paper on mathematical theory of the reconstructed process (Nee *et al.*, 1994), which considers extant species phylogenies to infer speciation and extinction rates. That allowed evolutionary biologists to test the developed mathematical theories with modern phylogenies.

The Yule process does not describe real phylogenies well (Blum and François, 2006), and species diversification models require more complex elaborations (Caron and Pie, 2020) and significant effort to satisfy both sensible biological and mathematical properties (Popovic, 2004). Biologically we would like to include many potential factors in our model taking into account the complexity of evolutionary processes, mathematically we would like to include chaotic dynamics well described by randomness or stochastic differential equations and statistically we would like to preserve identifiability. Since Nee's theory was developed, a large number of SDMs have been designed and tested using real

phylogenies. The utility of species diversification models resides in the option to quantify the relationship of potential covariates that could be related to species diversification processes defined by speciation and extinction rates,

$$\lambda_{t,s|\beta} = g_1\left(\sum_i^p \beta_{1i}\, v_{si}\right) \qquad \text{and} \qquad \mu_{t,s|\beta} = g_2\left(\sum_i^p \beta_{2i}\, v_{si}\right) \tag{1.1}$$

where a new species emerges from species $s$ at time $t$ with a speciation rate $\lambda_{t,s|\beta}$, which might be an arbitrarily (continous) function $g_1$ of a linear combination of, potentially, species-specific covariates $\{v_{s1}, ..., v_{sp}\}$, and species can become extinct with an extinction rate $\mu_{t,s|\beta}$. Phylogenetic trees are the result of these multiple events of speciation and extinction.

For example, various SDMs have been developed to test if diversification rates are related to the age of the species (Hagen *et al.*, 2015), to paleo-enviromental changes (Descombes *et al.*, 2018), to geographic patterns (Goldberg *et al.*, 2011), to time and space (Silvestro *et al.*, 2011), to a specific time dependency with key role in mass extinctions (Höhna, 2015), to species characters (Maddison *et al.*, 2007; Beaulieu and O'Meara, 2016; Herrera-Alsina *et al.*, 2019), to ecological fitness (Rasmussen and Stadler, 2019), to latitude (Schluter, 2016), to overall species diversity (Condamine *et al.*, 2019; Etienne *et al.*, 2012a), or whether speciation times are protracted (Lambert *et al.*, 2015; Etienne *et al.*, 2014), just to name a few. These individual models have created a kaleidoscopic view of the species diversification process, although their application in real phylogenies has been criticised (Rabosky, 2010, 2016), especially when estimating extinction rates from extant species phylogenies.

The statistical methods of this thesis can be used to test potentially *any* of the above-mentioned scenarios. We seek a unified methodology that considers complex interactions. Still, given that the list of possible factors affecting biodiversity is endless, we decided to focus all our illustrations on a specific class of models, the diversity-dependent (DD) diversification models. These models are relevant, because diversity can act as a proxy for many other ecological interactions.

### 1.1.1. DIVERSITY-DEPENDENT DIVERSIFICATION MODELS AND THE EFFECT OF ECOLOGICAL INTERACTIONS ON MACROEVOLUTIONARY PROCESSES

The presence of ecological limits to macroevolutionary processes has been hotly debated (Harmon and Harrison, 2015; Rabosky and Hurlbert, 2015). The simple and intuitive idea underlying models of diversity-dependent diversification is that speciation declines as diversity increases because the number of niches available to speciate into will decrease as more niches become occupied. This is often translated into a linear diversity-dependence as follows:

$$\lambda_{t;\beta} = \lambda_0 - \beta_N n_t; \qquad \mu_{t;\beta} = \mu_0, \qquad \lambda_0 > 0, \beta_N > 0, \mu_0 > 0 \tag{1.2}$$

where $\lambda_{t;\beta}$ is the individual speciation rate, $n_t$ the number of extant species and $\mu_{t,\beta}$ represents the individual extinction rate at time $t$, while $\lambda_0, \mu_0, \beta_N$ are parameters representing the initial speciation rate, the initial extinction rate and the decreasing slope

of speciation rate per species respectively. In this model the quantity $K' = \lambda_0/\beta_N$ is the value for which the size $n_t$ of a clade reaches a limit and cannot expand further as $\lambda_{t;\beta} = 0$. The carrying capacity $K$ is the value of diversity where the speciation and extinction rates equal one another, and this is related to $K'$ via $K = (\lambda_0 - \mu_0)K'/\lambda_0$. Thus, it is usually assumed that $\lambda_0$ is positive and $\beta_N$ is either negative or zero according to the idea that *the more species there are the less room there is for new species* (Etienne *et al.*, 2012b).

This model, even though it is currently widely used (Condamine *et al.*, 2019), is a simplification of the role that diversity plays in diversification. In the first place, it assumes that ecological limits to diversification are fixed in time, which is often not realistic (Marshall and Quental, 2016). Secondly, it assumes that diversity always has a negative effect on speciation, and hence that speciation rates cannot become larger than the initial speciation rate. Third, it uses species richness as a proxy for diversity, while in the ecological literature other measurements of diversity have been pointed out to be more representative (Magurran, 2013; Chao *et al.*, 2014). Fourth, it assumes that all species have the same probability of speciating, ignoring that species might occupy different ecological niches depending on their similarities and differences. Current inference models for diversity-dependent diversification, however, rely on these simplifications. While they can handle increasing rates of speciation as diversity increases or time-dependent diversification rates, they cannot deal with other measures of diversity or with differential rates between lineages, although some recent progress has been made (Laudanno *et al.*, 2020a). In this thesis I will develop methodology to enable inference under such more complex models of diversity-dependent species diversification.

The main component in the generalisations presented in this thesis is the incorporation of phylogenetic variation contained in clades of species. Thus, in Chapters 2 and 3 we consider the phylogenetic diversity of the clade, as a function of time, in the phylogenetic diversity-dependent diversification model:

$$\lambda_{t|\theta} = \lambda_0 + \beta_n n_t + \beta_p \frac{p_t - t}{n_t}; \qquad \mu_{t|\beta} = \mu_0 \tag{1.3}$$

where $p_t$ is the phylogenetic diversity at time $t$. The quantity $\frac{p_t - t}{n_t}$ corresponds to the *phylogenetic diversity per species* at time $t$. Note that by subtracting $t$ the phylogenetic diversity per species in the case of a single species stays 0, as required.

In Chapter 4 we focus on lineage-dependent diversification models where diversification rates depend on the phylogenetic uniqueness of species:

$$\lambda_{t,s|\theta} = \lambda_0 + \beta_N n_t + \beta_P \frac{(P_{t,s} - \bar{P}_t)}{n_t}; \qquad \mu_{t;\beta} = \mu_0 \tag{1.4}$$

where $P_{t,s}$ is a measure of the phylogenetic uniqueness of species $s$ at time $t$ and $\bar{P}_t = \sum_s P_{t,s}/n_t$ is a measure of the overall phylogenetic diversity in the clade.

### 1.1.2. EXAMPLE

In Figure 1.2, we consider an example phylogenetic tree. The plots underneath describe different quantities included in the evolutionary process: species richness, global

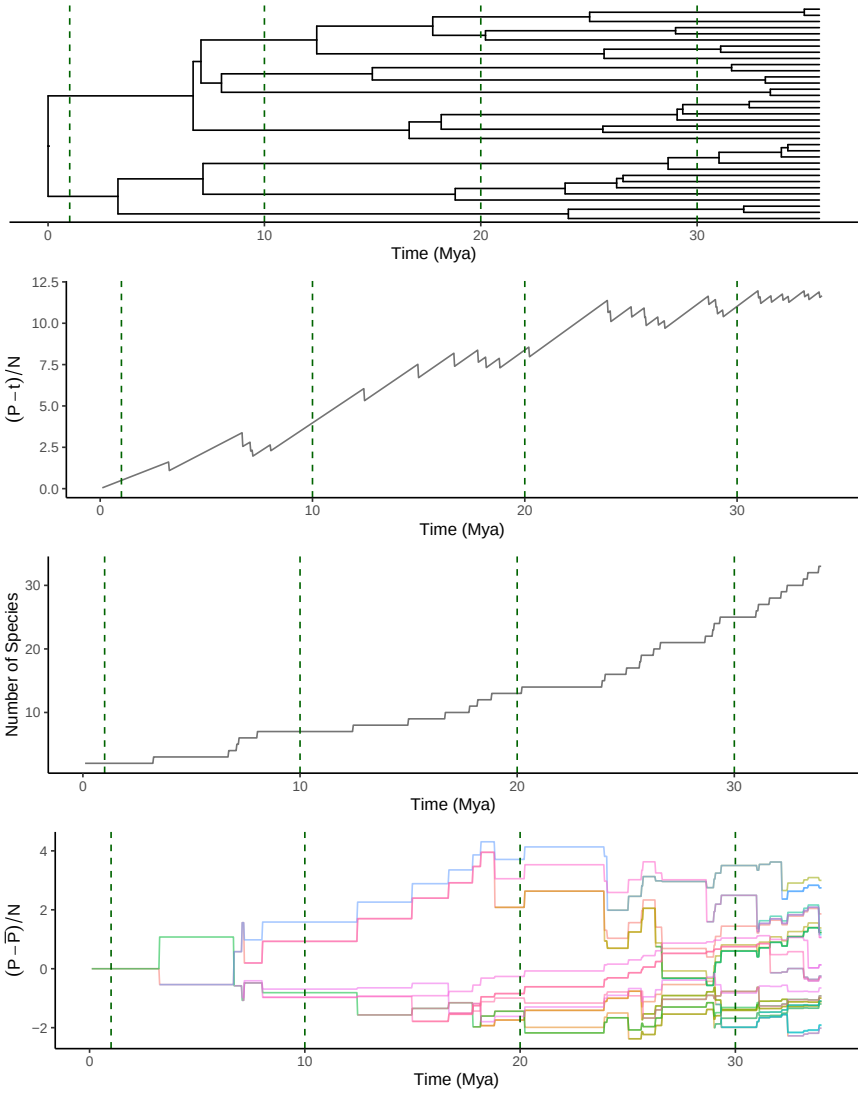



**Figure 1.2 |** Visualization of various types of diversity. At the top we see a phylogenetic tree of extant species, corresponding to the clade *Bucconidae*. The second plot is the global phylogenetic diversity per species through time. The third plot corresponds to the number of lineages through time. The final plot shows the mean pairwise phylogenetic diversity and the normalised pairwise phylogenetic diversity per species through time.

phylogenetic diversity, and mean pairwise phylogenetic diversity, respectively. We calculate the $P$ matrix of pairwise phylogenetic diversity for four different times $t$. At time $t = 1$

the tree has only two species, so the P matrix is

$$P(1) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

When new species emerge, the $P$ matrix increases in dimensions. At $t = 10$ the pairwise phylogenetic distances can be summarised with the matrix

$$P(10) = \begin{bmatrix} 0 & 10 & 6.8 & 10 & 10 & 2.8 & 10 \\ 10 & 0 & 10 & 3.3 & 2.9 & 10 & 2.9 \\ 6.8 & 10 & 0 & 10 & 10 & 6.8 & 10 \\ 10 & 3.3 & 10 & 0 & 3.3 & 10 & 3.3 \\ 10 & 2.9 & 10 & 3.3 & 0 & 10 & 2 \\ 2.8 & 10 & 6.8 & 10 & 10 & 0 & 10 \\ 10 & 2.9 & 10 & 3.3 & 2 & 10 & 0 \end{bmatrix}$$

while at time $t = 20$ the matrix is

$$P(20) = \begin{bmatrix} 0 & 20 & 16.8 & 20 & 20 & 12.8 & 20 & 20 & 20 & 20 & 20 & 20 & 1.2 \\ 20 & 0 & 20 & 13.3 & 12.9 & 20 & 12.9 & 7.6 & 12.9 & 13.3 & 2.2 & 13.3 & 20 \\ 16.8 & 20 & 0 & 20 & 20 & 16.8 & 20 & 20 & 20 & 20 & 20 & 20 & 16.8 \\ 20 & 13.3 & 20 & 0 & 13.3 & 20 & 13.3 & 13.3 & 13.3 & 3.3 & 13.3 & 3.3 & 20 \\ 20 & 12.9 & 20 & 13.3 & 0 & 20 & 12 & 12.9 & 12 & 13.3 & 12.9 & 13.3 & 20 \\ 12.8 & 20 & 16.8 & 20 & 20 & 0 & 20 & 20 & 20 & 20 & 20 & 20 & 12.8 \\ 20 & 12.9 & 20 & 13.3 & 12 & 20 & 0 & 12.9 & 5 & 13.3 & 12.9 & 13.3 & 20 \\ 20 & 7.6 & 20 & 13.3 & 12.9 & 20 & 12.9 & 0 & 12.9 & 13.3 & 7.6 & 13.3 & 20 \\ 20 & 12.9 & 20 & 13.3 & 12 & 20 & 5 & 12.9 & 0 & 13.3 & 12.9 & 13.3 & 20 \\ 20 & 13.3 & 20 & 3.3 & 13.3 & 20 & 13.3 & 13.3 & 13.3 & 0 & 13.3 & 1.8 & 20 \\ 20 & 2.2 & 20 & 13.3 & 12.9 & 20 & 12.9 & 7.6 & 12.9 & 13.3 & 0 & 13.3 & 20 \\ 20 & 13.3 & 20 & 3.3 & 13.3 & 20 & 13.3 & 13.3 & 13.3 & 1.8 & 13.3 & 0 & 20 \\ 1.2 & 20 & 16.8 & 20 & 20 & 12.8 & 20 & 20 & 20 & 20 & 20 & 20 & 0 \end{bmatrix}$$

and at time $t = 30$ the matrix has 25 rows and columns.

$P(30) =$

| | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 30 | 26.8 | 30 | 30 | 22.8 | 30 | 30 | 30 | 30 | 30 | 30 | 11.2 | 30 | 11.2 | 26.8 | 30 | 30 | 30 | 11.2 | 11.2 | 22.8 | 30 | 30 | 30 |
| 30 | 0 | 30 | 23.3 | 22.9 | 30 | 22.9 | 17.6 | 22.9 | 23.3 | 12.2 | 23.3 | 30 | 9.8 | 30 | 30 | 12.2 | 23.3 | 17.6 | 30 | 30 | 30 | 9.8 | 25.3 | 25.3 |
| 26.8 | 30 | 0 | 30 | 30 | 26.8 | 30 | 30 | 30 | 30 | 30 | 30 | 26.8 | 30 | 26.8 | 6 | 30 | 30 | 30 | 26.8 | 26.8 | 26.8 | 30 | 30 | 30 |
| 30 | 23.3 | 30 | 0 | 23.3 | 30 | 23.3 | 25.3 | 23.3 | 13.3 | 23.3 | 13.3 | 30 | 25.3 | 30 | 30 | 23.3 | 13.3 | 25.3 | 30 | 30 | 30 | 25.3 | 15.3 | 15.3 |
| 30 | 22.9 | 30 | 23.3 | 0 | 30 | 22 | 22.9 | 22 | 23.3 | 22.9 | 23.3 | 30 | 22.9 | 30 | 30 | 22.9 | 23.3 | 22.9 | 30 | 30 | 30 | 22.9 | 25.3 | 25.3 |
| 22.8 | 30 | 26.8 | 30 | 30 | 0 | 30 | 30 | 30 | 30 | 30 | 30 | 22.8 | 30 | 22.8 | 26.8 | 30 | 30 | 30 | 22.8 | 22.8 | 1.3 | 30 | 30 | 30 |
| 30 | 22.9 | 30 | 23.3 | 22 | 30 | 0 | 22.9 | 15 | 23.3 | 22.9 | 23.3 | 30 | 22.9 | 30 | 30 | 22.9 | 23.3 | 22.9 | 30 | 30 | 30 | 22.9 | 25.3 | 25.3 |
| 30 | 17.6 | 30 | 23.3 | 22.9 | 30 | 22.9 | 0 | 22.9 | 23.3 | 17.6 | 23.3 | 30 | 17.6 | 30 | 30 | 17.6 | 23.3 | 4.3 | 30 | 30 | 30 | 17.6 | 25.3 | 25.3 |
| 30 | 22.9 | 30 | 23.3 | 22 | 30 | 15 | 22.9 | 0 | 23.3 | 22.9 | 23.3 | 30 | 22.9 | 30 | 30 | 22.9 | 23.3 | 22.9 | 30 | 30 | 30 | 22.9 | 25.3 | 25.3 |
| 30 | 23.3 | 30 | 13.3 | 23.3 | 30 | 23.3 | 25.3 | 23.3 | 0 | 23.3 | 11.8 | 30 | 23.3 | 30 | 30 | 23.3 | 11.8 | 23.3 | 30 | 30 | 30 | 23.3 | 0.9 | 0.7 |
| 30 | 12.2 | 30 | 23.3 | 22.9 | 30 | 22.9 | 17.6 | 22.9 | 25.3 | 0 | 23.3 | 30 | 12.2 | 30 | 30 | 5 | 23.3 | 17.6 | 30 | 30 | 30 | 12.2 | 25.3 | 25.3 |
| 30 | 23.3 | 30 | 13.3 | 23.3 | 30 | 25.3 | 25.3 | 23.3 | 11.8 | 23.3 | 0 | 30 | 25.3 | 30 | 30 | 23.3 | 4.4 | 23.3 | 30 | 30 | 30 | 23.3 | 11.8 | 11.8 |
| 11.2 | 30 | 26.8 | 30 | 30 | 22.8 | 30 | 30 | 30 | 30 | 30 | 30 | 0 | 30 | 6.1 | 26.8 | 30 | 30 | 30 | 3.7 | 3.7 | 22.8 | 30 | 30 | 30 |
| 30 | 9.8 | 30 | 23.3 | 22.9 | 30 | 22.9 | 17.6 | 22.9 | 23.3 | 12.2 | 23.3 | 30 | 0 | 30 | 30 | 12.2 | 23.3 | 17.6 | 30 | 30 | 30 | 1 | 25.3 | 25.3 |
| 11.2 | 30 | 26.8 | 30 | 30 | 22.8 | 30 | 30 | 30 | 30 | 30 | 30 | 6.1 | 30 | 0 | 26.8 | 30 | 30 | 30 | 6.1 | 6.1 | 22.8 | 30 | 30 | 30 |
| 26.8 | 30 | 6 | 30 | 30 | 26.8 | 30 | 30 | 30 | 30 | 30 | 30 | 26.8 | 30 | 26.8 | 0 | 30 | 30 | 30 | 26.8 | 26.8 | 26.8 | 30 | 30 | 30 |
| 30 | 12.2 | 30 | 23.3 | 22.9 | 30 | 22.9 | 17.6 | 22.9 | 23.3 | 5 | 23.3 | 30 | 12.2 | 30 | 30 | 0 | 23.3 | 17.6 | 30 | 30 | 30 | 12.2 | 25.3 | 25.3 |
| 30 | 23.3 | 30 | 13.3 | 23.3 | 30 | 23.3 | 25.3 | 23.3 | 11.8 | 23.3 | 4.4 | 30 | 23.3 | 30 | 30 | 23.3 | 0 | 23.3 | 30 | 30 | 30 | 23.3 | 11.8 | 11.8 |
| 30 | 17.6 | 30 | 23.3 | 22.9 | 30 | 22.9 | 4.3 | 22.9 | 23.3 | 17.6 | 23.3 | 30 | 17.6 | 30 | 30 | 17.6 | 23.3 | 0 | 30 | 30 | 30 | 17.6 | 25.3 | 25.3 |
| 11.2 | 30 | 26.8 | 30 | 30 | 22.8 | 30 | 30 | 30 | 30 | 30 | 30 | 3.7 | 30 | 6.1 | 26.8 | 30 | 30 | 30 | 0 | 3.4 | 22.8 | 30 | 30 | 30 |
| 11.2 | 30 | 26.8 | 30 | 30 | 22.8 | 30 | 30 | 30 | 30 | 30 | 30 | 3.7 | 30 | 6.1 | 26.8 | 30 | 30 | 30 | 3.4 | 0 | 22.8 | 30 | 30 | 30 |
| 22.8 | 30 | 26.8 | 30 | 30 | 1.3 | 30 | 30 | 30 | 30 | 30 | 30 | 22.8 | 30 | 22.8 | 26.8 | 30 | 30 | 30 | 22.8 | 22.8 | 0 | 30 | 30 | 30 |
| 30 | 9.8 | 30 | 23.3 | 22.9 | 30 | 22.9 | 17.6 | 22.9 | 23.3 | 12.2 | 23.3 | 30 | 1 | 30 | 30 | 12.2 | 23.3 | 17.6 | 30 | 30 | 30 | 0 | 25.3 | 25.3 |
| 30 | 23.3 | 30 | 13.3 | 23.3 | 30 | 23.3 | 25.3 | 23.3 | 0.9 | 23.3 | 11.8 | 30 | 23.3 | 30 | 30 | 23.3 | 11.8 | 23.3 | 30 | 30 | 30 | 23.3 | 0 | 0.9 |
| 30 | 23.3 | 30 | 13.3 | 23.3 | 30 | 23.3 | 25.3 | 23.3 | 0.7 | 23.3 | 11.8 | 30 | 25.3 | 30 | 30 | 23.3 | 11.8 | 23.3 | 30 | 30 | 30 | 23.3 | 0.9 | 0 |

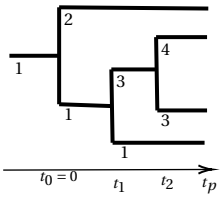
## 1.2. THE MODE AND TEMPO OF DIVERSIFICATION PROCESSES

Mathematically, phylogenetic trees have two components: branching times and topology (Ragan, 2009). In Figure 1.3 we see a tree and matrix representation of an extant species tree (i.e. an ultrametric tree) and a full tree containing extinctions. The first column represents the branching times while the next two columns represent the topology. These two are the mathematical expressions of the mode and tempo of a dicersification process.

While the timing (or tempo) in the tree can be represented by the total sum of rates of the system, and hence, lineage-independent diversification models could capture such

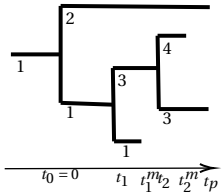
behaviour, the topology of the tree and hence its balance affect the diversification rates of different lineages differently (Heard, 1996), requiring an extra level of complexity (Savage, 1983) than is available in most of the SDMs for which currently inference methods exist.

Ultrametric trees

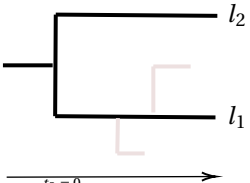



$$\begin{bmatrix} Brts & Child & Parent \\ 0 & 2 & 1 \\ t_1 & 3 & 1 \\ t_2 & 4 & 3 \\ t_p & & \end{bmatrix}$$

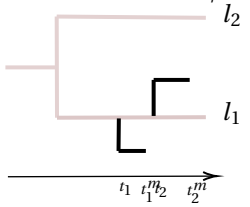
Non-ultrametric trees




$$\begin{bmatrix} Brts & Child & Parent & Extinction \\ 0 & 2 & 1 & \\ t_1 & 3 & -1 & t_1^m \\ t_2 & -4 & 3 & t_2^m \\ t_p & & & \end{bmatrix}$$




$$\begin{bmatrix} Brts & Child & Parent \\ 0 & 2 & 1 \\ t_p & & \end{bmatrix}$$




$$\begin{bmatrix} Brts & Child & Parent & Extinction \\ t_1 & 3 & -1 & t_1^m \\ t_2 & -4 & 3 & t_2^m \end{bmatrix}$$

**Figure 1.3 |** Phylogenetic trees represented in a diagram and the corresponding matrix notation

In Figure 1.4 we show the mode or topology and tempo characterised by the Colless and Gamma indices, respectively, for 218 real phylogenies. The first observation in the plot is that real phylogenies are unbalanced, given that the Colless index is always different from zero. The Gamma index values show variability among clades, from negative to positive, showing that some phylogenies show slowdowns in lineage accumulating through time, whereas others speed up. The families of the clades do not show different tendencies among them. Various indices have been developed to capture both balance and timing (Mir *et al.*, 2013). In this thesis, we do not compare them, but we use two of the

most common ones instead. However, all analyses can be easily repeated with alternative indices.

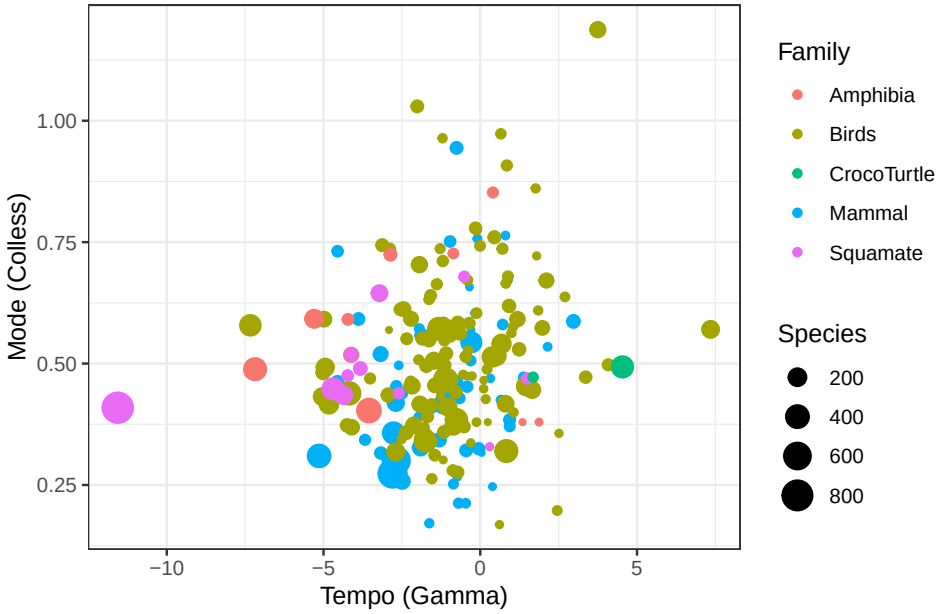



**Figure 1.4 |** Distribution of Colless and Gamma index for 5 families describing Mode and Tempo respectively.

## 1.3. STATISTICAL METHODOLOGIES

This thesis deals with the development of inference techniques for the mathematical models describing the species diversification process on the basis of extant phylogenetic trees. The inference procedure we propose relies on a number of statistical techniques that we describe in this section.

### 1.3.1. THE LIKELIHOOD APPROACH

By assuming a functional form of the probability distributions of the species diversification process, statistical testing (Casella and Berger, 2002; Pfanzagl, 2011) of a wide variety of hypotheses in macroevolution is possible. The maximum likelihood (ML) method consists of calculating the parameters $\theta$ that maximise the probability of observing the data $y$,

$$\hat{\theta} = \arg\max_{\theta} f_M(y|\theta).$$

Multiples challenges arise when performing ML estimation in the context of phylogenetic trees, for example because covariates are typically only observed at the present but they do contribute to speciation and extinction rates throughout the whole evolutionary process.

A large number of packages have been developed for specific species diversification processes, but there is no unified framework that considers all of them. This thesis is a step towards such a framework.

Current methods to compute the likelihood of a given SDM attempt to solve the so-called *master equations* or *Kolmogorov equations* (Carmona and Delarue, 2014). For each species diversification model, a new likelihood needs to be calculated, if at all possible. In this thesis, we implemented a different approach, that does not require solving any master equations.

### 1.3.2. EM ALGORITHM

In Chapters 2 and 3 we make use of a novel implementation of the EM algorithm for inference of SDM parameters. The EM algorithm is an iterative procedure that optimises a loglikelihood by calculating the expected loglikelihood given some parameters (E-step) and then maximising it (M-step) providing a new set of parameters for the next iteration. The EM algorithm is proven to increase the likelihood in every iteration and converge to the ML estimate for convex likelihoods.

### 1.3.3. MONTE-CARLO

The Monte-Carlo algorithm is a numerical method for integration that cannot be calculated analytically or numerically by standard methods. By sampling $n$ realisations, an approximation of the integral of a function $f$ can be calculated as follows,

$$\int_{x\in\mathscr{X}} f(x)\, dx \approx \frac{c}{n} \sum_{x_i \sim U(\mathscr{X})} f(x_i), \tag{1.5}$$

where the $x_i$ are $n$ uniform draws from $\mathscr{X}$ and $c = \int_{x\in\mathscr{X}} dx$. In the context presented here, $x$ is the unobserved part of the tree, typically the species that did not make it to the present, and hence for computing the likelihood of an extant-species phylogeny, we need to integrate over all possible extinction patterns that could have given rise to the current extant tree. When $n \to \infty$ the approximation converges to an equality. However, sampling uniformly on $\mathscr{X}$ is not always possible, or it can be particularly inefficient, if for many values $x \in \mathscr{X}$ the function $f$ contributes negligibly to the integral, $f(x) \approx 0$. To deal with this problem, we use an importance sampling technique.

### 1.3.4. IMPORTANCE SAMPLING AND DATA AUGMENTATION

Importance sampling is a statistical technique that can be used when sampling from the desired distribution is difficult. It consists of sampling from an alternative distribution that contains the support of the desired distribution and correcting for it by the ratio of the probability $f(x)$ according to the true distribution by the probability of the realisation $x$ according to the alternative sampling distribution. Thus, we replace equation 1.5 by

$$\int_{x\in\mathscr{X}} f(x)dx \approx \frac{1}{n} \sum_{x_i \sim g} \frac{f(x_i)}{g(x_i)} \tag{1.6}$$

Observe that choosing $g = f$ leads to Eq. 1.5. A vital component of this method is to find a "good" importance sampler $g$. Given that in the current context we need to sample the extinct species on top of the extant or observed species of the tree, we introduce a data augmentation algorithm that samples full trees that are in agreement with the observed trees. In Chapter 2 we use a uniform importance sampler, which is not efficient in the sense that it does not simulate trees with significant values of $f$, given that is ignorant about the true process and parameters, but it is a good starting point for further comparisons as well as simple to implement and interpret. In Chapter 3 we develop a sophisticated data augmentation algorithm which allows us to implement the method for large trees.

The data augmentation algorithms developed in this thesis are not only useful for our EM algorithm but also for a wide variety of parameter estimation methods including Bayesian methods or stochastic gradient descent approaches.

### 1.3.5. STOCHASTIC GRADIENT DESCENT METHOD

In Chapter 4 we propose an alternative to the EM algorithm by maximising the likelihood function using a stochastic gradient descent method which is also an iterative procedure, but now each iteration calculates the next parameter value using

$$\theta_i = \theta_{i-1} - \eta\, G(\theta)$$

where $\eta$ is a step size, also known as the learning rate in the machine learning literature, and $G$ is the gradient of the likelihood function. The gradient of the observed likelihood is defined as

$$G(\theta) = \frac{\partial}{\partial\theta}\int_{x\in\mathscr{X}} f(x,y|\theta)\,dx$$

Here we use again our data augmentation algorithm developed in Chapter 3 for an unbiased estimate of the gradient as the direct calculation of this expectation is, again, not possible. Thus we estimate the gradient $G$ by its Monte Carlo approximation.

Comparison and similarities between the EM algorithm and gradients methods have been studied (Xu and Jordan, 1996). In this thesis we make use of both of them but do not necessarily compare them as they are used in different contexts: Chapters 2 and 3, which employ the EM, are implementations of lineage-independent diversification models while Chapter 4, where we exploit the gradient descent method, contains the implementation of lineage-dependent diversification models.

### 1.3.6. GENERALISED ADDITIVE MODELS

Throughout this thesis we will come across expressions of the likelihood that are complex functions of the parameters. Sometimes, we resolve this issue by means of Monte Carlo sampling, but in other cases we try to approximate the functional form through the theory of generalised additive models (GAM). GAMs are smooth functions that can approximate any continuous function. By fitting linear combinations of polynomials, typically cubic splines, computationally efficient approximations can be obtained.

We argue that this theory can be applied as well to estimate a general class of species diversification models by approximating the likelihood function using GAMs, at least for models with few parameters. In this thesis, we use the GAM approach to calculate a conditioning probability as a function of the parameters. Chapter 5 is dedicated to developing a method to calculate the probability of any conditioning event for any SDM that can be recorded by simulation.

## 1.4. THE CONDITIONED EVOLUTIONARY PROCESS

The macroevolutionary processes as we see it now is undoubtedly conditioned to, at least, the fact that we indeed observe it, i.e., that the process survives to the present. Other conditioning arguments can be found in the literature to incorporate in the likelihood function of the species diversification process. Suggested conditioning events are the survival of the tree given a particular crown or stem age, or the number of tips (Gernhard, 2008) or both (Etienne *et al.*, 2016; Stadler, 2013). However, expressions for calculating these conditioning probabilities as a function of the SDM parameters have only been developed for specific species diversification models. In Chapter 5 of this thesis, we develop a novel implementation for conditioning, where we make use of a combination of simulations and GAM smoothing to estimate the probability of any condition that can be used to calculate the conditioned likelihood of the process.

## 1.5. MODEL SELECTION

Model selection is a standard topic in statistical inference. Information criteria such as AIC or BIC are well known and used in many statistical applications to selecting the "best" model from a set of them (Wit *et al.*, 2012). These techniques are used when a likelihood is available. In the context of species diversification models and phylogenetic trees, alternative methods have been used to compare the goodness-of-fit of a model to real data. Simple statistics involve the number of tips or the gamma statistic, which take into account the timing distribution of the phylogenetic tree. However, they do not provide a strong tool for model selection. A more sophisticated statistic that captures the whole evolutionary process is the *lineage-through-time* (LTT) statistic (Janzen *et al.*, 2015). The LTT statistic is defined as the integral of the absolute difference between two phylogenetic trees

$$LTT(1,2) = \int_{t_0}^{t_p} |N_1(t) - N_2(t)| dt$$

where $N_i(t)$ is the number of lineages in the tree $i$ at time $t$. In Figure 1.5 we see the area between two LTT plots corresponding to two different trees in blue. This value can be used to compare two trees but also to assess how similar trees simulated from a specific model are to the observed tree. Moreover, it can be used to estimate parameters using a likelihood-free method such as ABC inference (Janzen *et al.*, 2015). The LTT plot is a curve that does not consider the topology of the tree. Alternatively, we propose the *phylodiversity-through-time* (PTT) statistic. This quantity does consider the topology of the tree by integrating the difference between the phylodiversities of two trees through

time,

$$PTT(1,2) = \int_{t_0}^{t_p} |P_1(t) - P_2(t)| dt.$$

In the literature this has not been studied yet. In the bottom panel of Figure 1.5, we visualise the difference between phylodiversities at each time point of the two trees.

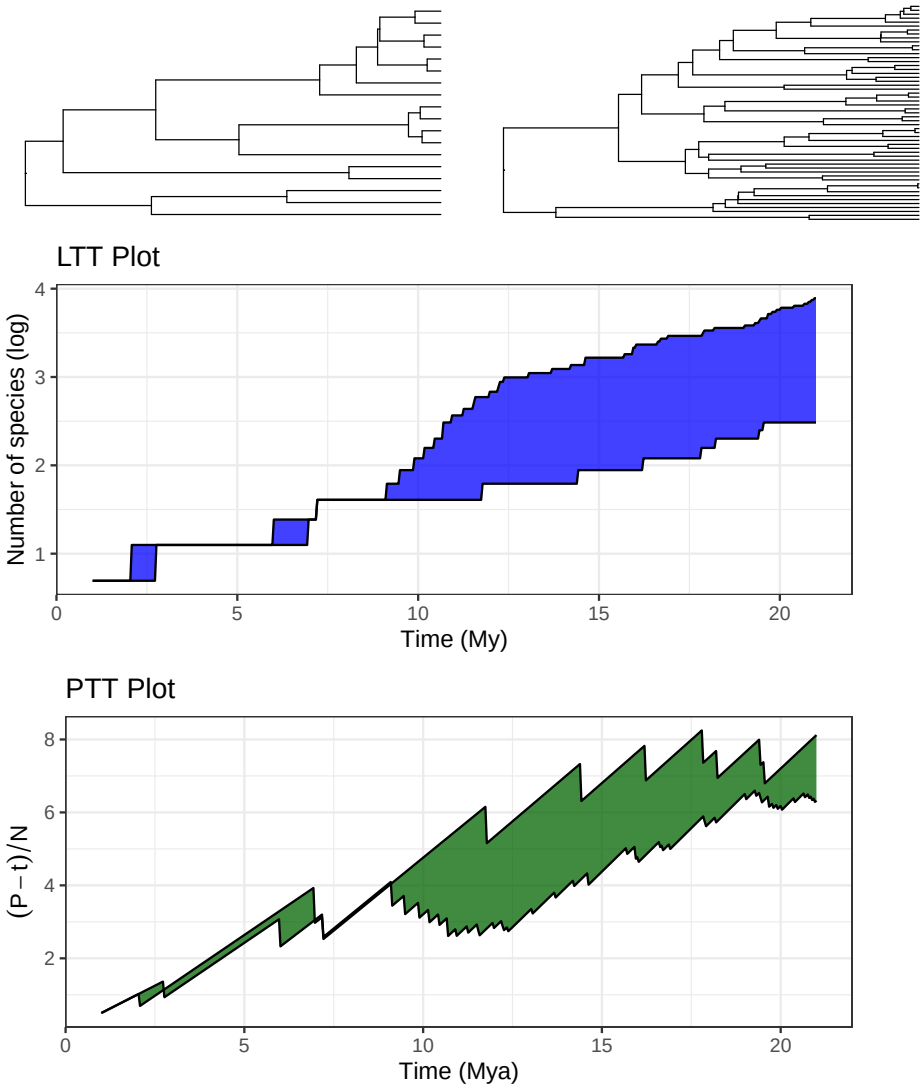



**Figure 1.5 |** Comparison among two phylogenetic trees. Number of lineages and Phylogenetic diversity through time. The area represents the distance between the trees.

## 1.6. OUTLINE OF THE THESIS

This thesis lies on the intersection of evolutionary biology and statistical network science, providing a systematic development in both areas. Throughout the thesis we increase the generality of our approach to modelling phylogenetic trees in order to achieve our two main objectives. On the one hand, we aim to develop a methodology to perform statistical inference in phylogenetics and macroevolution, allowing for testing of a wide variety of species diversification dynamics hypotheses. On the other hand, we develop generalisations of diversity-dependent diversification models, because they are ideal systems to consider, study and quantify the effect of ecological limits and species interactions on species diversification processes.

In Chapter 2, we develop an *Monte Carlo Expectation-Maximization* (MCEM) type of algorithm in the context of phylogenetic trees, which in combination with a data augmentation algorithm is a powerful tool for flexible statistical inference and species evolution modelling. In this chapter, we provide a simple data augmentation scheme as a basis for future comparisons, while achieving a fast general algorithm for small phylogenetic trees.

In Chapter 3, we present an efficient and elegant data augmentation algorithm (DAA) for reconstructed phylogenetic trees. The new DAA together with the MCEM method developed in Chapter 2, make it possible to perform statistical inference in medium-size phylogenetic trees. The method is called *emphasis*, which stands for Expectation-Maximization in PHylogenetic Analysis with Simulations and Importance Sampling. In this chapter we also generalise diversity-dependent diversification models including phylodiversity, an essential type of diversity that considers the genetic distance among species and thus provides another dimension to the niche filling argument that previous diversity-dependent diversification models were aiming to capture.

In Chapter 4, we generalise diversity-dependent diversification models even further by allowing lineage-specific evolutionary advantage with lineage-dependent diversification models. Moreover, we develop a stochastic gradient descent algorithm to incorporate in the statistical toolkit for parameter estimation, making use of the data augmentation algorithm developed in Chapter 3. We provide a model that is flexible and capable of capturing a large variety of topologies.

In Chapters 2-4, the phylogenetic space considered were all possible phylogenetic trees. However, the very fact that any inference of data only makes sense for a surviving clade, there are implicit assumptions, e.g., about the existence of the clade. The practice of conditioning the observed data is a common practice in phylogenetic analysis. Conditioning means that only a certain part of outcome space is considered feasible for the observed phylogeny. Although various types of conditioning are used in phylogenetic analysis, the most common, and most intuitive, one is that the observed phylogeny is not empty. This affects the likelihood function and maximum likelihood estimates. Explicit conditioning formulas are elusive, but we dedicate Chapter 5 exclusively to developing a method that is able to approximate the probability of any condition of a process under any species diversification model.

Chapter 6 presents a collection of considerations that have come up over the course of this PhD project. It presents the limitations of the methods presented in this thesis

as well as considerations for improving some of these methods. Ways of dealing with incomplete clades as well as with possible extensions to the various models are discussed. Furthermore, it considers how the methods used in this thesis can be used outside the field of evolutionary biology.

# 2

# INTRODUCING A GENERAL CLASS OF SPECIES DIVERSIFICATION MODELS FOR PHYLOGENETIC TREES

*I would build my world which while I lived, would be in agreement with all the worlds.*

Frida Kahlo

Richter F, Haegeman B, Etienne RS, Wit EC. Introducing a general class of species diversification models for phylogenetic trees. Statistica Neerlandica. 2020;1–14. https://doi.org/10.1111/stan.12205

## ABSTRACT

*Phylogenetic trees are types of networks that describe the temporal relationship between individuals, species or other units that are subject to evolutionary diversification. Many phylogenetic trees are constructed from molecular data which is often only available for extant species, and hence they lack all or some of the branches that did not make it into the present. This feature makes inference on the diversification process challenging. For relatively simple diversification models analytical or numerical methods to compute the likelihood exist, but these do not work for more realistic models in which the likelihood depends on properties of the missing lineages. In this paper we study a general class of species diversification models, and we provide an expectation-maximization framework in combination with a uniform sampling scheme to perform maximum likelihood estimation of the parameters of the diversification process.*

## 2.1. INTRODUCTION

Evolutionary relationships of species are commonly described by phylogenetic trees or, in more general scenarios, by phylogenetic networks (Ragan, 2009). A phylogenetic tree is a hypothesis on how species or other biological units have diversified over time. It is usually described by a binary tree whose nodes are ordered in time. Phylogenetic relationships can be inferred from a variety of sources such as morphology and behaviors of species, biochemical pathways, DNA and protein sequences (Lemey *et al.*, 2009), both from extant, i.e., living species or from extinct species through ancient DNA or the fossil record. However, data on extinct species is often incomplete and only accurate molecular phylogenies of extant species are available. In this manuscript we consider such phylogenetic trees as primary observations. Even though they lack extinct lineages, they are believed to contain information on how species diversified and hence they have been used to answer fundamental questions, such as "does diversity affect diversification?" (Etienne *et al.*, 2012b; Cornell, 2013), "what is the effect of environmental and ecological interactions on evolutionary dynamics?" (Ezard *et al.*, 2011; Barraclough, 2015; Lewitus and Morlon, 2017), "how does biodiversity vary spatially?" (Goldberg *et al.*, 2011; Mittelbach *et al.*, 2007), and "what traits play a key role in species diversification?" (Lynch, 2009; Paradis, 2005; FitzJohn *et al.*, 2009), to name just a few.

To help to answer these questions specific mathematical models have been developed that can infer various parameters from phylogenetic diversification pattern (Morlon, 2014). Most current approaches have started to use likelihood based methods to perform inference on phylogenetic trees (Etienne *et al.*, 2012b; FitzJohn *et al.*, 2009; Ricklefs, 2007; Stadler, 2011, e.g.). Although statistically principled, in each of these models a new method to compute the likelihood needs to be developed. These models often rely on describing the macro-evolutionary process by coupled ordinary differential equations — the so-called *master* or *Kolmogorov equations* — and these quickly become intractable as model complexity increases, particularly due to the lack of data on extinct species (Ricklefs, 2007; Höhna *et al.*, 2011).

Alternative ways to deal with Kolmogorov equations have been used since the 1950s in fields outside evolutionary biology. These methods have used point process theory (Serfozo, 1990; Daley and Vere-Jones, 2007), which does not solve Kolmogorov equations directly but employs Gillespie-type simulations that were introduced in the context of chemical reaction modelling (Gillespie, 1976, 1977). A single Gillespie simulation represents an exact sample from the probability mass function that is the solution of the system, thus allowing for stochastic optimization methods to maximize the likelihood (Tijms, 1994).

In this paper we present a first step for a general inference procedure of a general species diversification model. In section 2.2 we describe a general diversification process based on a generalized linear model description of a non-homogeneous point process. This model can be used to describe many alternative evolutionary hypotheses. In section 2.3 we introduce an expectation-maximization (EM) algorithm to optimize the likelihood under incomplete information, namely the extinct lineages. We present a data augmentation algorithm, involving stochastic simulation combined with an importance sampler, to perform the E step. We provide a proof-of-concept by comparing our inference with that

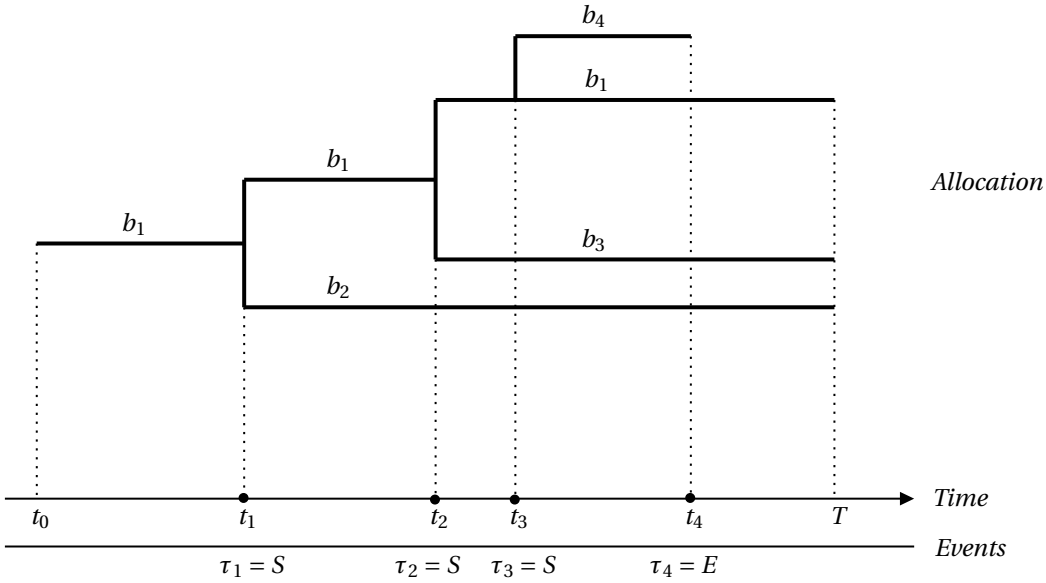



**Figure 2.1 |** Phylogenetic tree with four events: three speciation events and one extinction event. Each branch represents a species.

obtained using direct likelihood calculations. In section 2.4 we apply our method to the diversification of a small clade of Vangidae, consisting of a group of medium-sized birds living in Madagascar. Our aim is to discover whether the evolutionary record supports more the diversity dependence hypothesis (Etienne *et al.*, 2012b) or the phylodiversity hypothesis (Castillo *et al.*, 2010), for which no direct likelihood computation exists. Finally in Section 2.5 we provide directions for future extensions of the method that are needed to allow evolutionary biologists to routinely apply our approach to larger phylogenetic trees to study general diversification dynamics in a unified framework.

## 2.2. A GENERAL DIVERSIFICATION MODEL

We define a phylogenetic tree $x = (\tau, t, a)$ on a time interval $[0, T]$ as a functional object described by three components: a binary vector $\tau$ of event types (speciation or extinction), a vector of continuous event times $t$ and a network configuration object $a$, describing which species speciated or went extinct at each event time. We model the shape and structure of the tree by means of a collection of point processes, in this case, a set of dynamical non-homogeneous Poisson processes (NHPP) where speciation and extinction of species are random events that happen within a time interval $[0, T]$. Figure 2.1 shows an example of a phylogenetic tree with three speciation events and one extinction event.

In this paper, we assume that the process starts at time $t_0 = 0$ with a single species $b_1$. At this stage, the tree is subject to two Poisson processes: a potential speciation of species

$b_1$ and a potential extinction of species $b_1$. Both processes are assumed to have a waiting time with time-continuous rates $\lambda_{b_1}(t)$ and $\mu_{b_1}(t)$, respectively. In the time-homogeneous case, the waiting time for the first event to occur is therefore an exponential with rate $\lambda_{b_1} + \mu_{b_1}$. More generally (Daley and Vere-Jones, 2007), the probability density for the process $x$ to have a single species up to time $t_1$ and a speciation event exactly at time $t_1$ is given by

$$f(t_1) = \lambda_{b_1}(t_1) e^{-\int_{t_0}^{t_1} \lambda_{b_1}(t) + \mu_{b_1}(t) dt}. \tag{2.1}$$

If indeed a speciation occurs, the process continues with four NHPPs: two potential speciations and two potential extinctions. This is repeated until the present time $T$, unless the tree dies out before then. We consider a general scenario where at time $t$ each of the $N_t$ present species $b$ has its own speciation rate $\lambda_b(t)$ and extinction rate $\mu_b(t)$ defined as a linear function via link function $h$,

$$h(\lambda_b(t)) = \sum_{j=1}^{m} \beta_j c_{bjt}, \quad h(\mu_b(t)) = \sum_{j=1}^{m} \alpha_j c_{bjt}. \tag{2.2}$$

where $c_{bjt}$ is one of $j = 1, \ldots, m$ possible covariates of species $b$ at time $t$ affecting the speciation and/or extinction processes. Our entire process is therefore governed by the parameter set $\theta = \{\beta_1, \ldots, \beta_m, \alpha_1, \ldots, \alpha_m\}$. Typically we will consider the logarithmic link function $h = \log$, but equation (2.2) can be trivially modified by choosing for $h$ any monotonous increasing function that maps $(0, \infty)$ onto $\mathbb{R}$. The class of statistical models satisfying these specifications are an extension of the well-known generalized linear models (GLMs) (Dobson and Barnett, 2008).

This GLM extension to phylogenetic trees spans a very broad spectrum of possibilities for evolutionary biologists to test hypotheses and integrate their species diversification data. Diversification rates can be influenced by individual attributes, typically called *traits*, environmental factors, such as average temperature, by the composition of the diversifying clade itself or of its local ecological community. In the literature a range of models have been explored, where diversification rates are assumed to be constant (Nee *et al.*, 1994), change through time (Rabosky and Lovette, 2008), depend on diversity (Etienne *et al.*, 2012b), on individual traits (Paradis, 2005; Freckleton *et al.*, 2008) or other factors (Morlon, 2014). In order to test realistic models, we are interested in flexible rates that are able to change dynamically through all those factors simultaneously. For example, the speciation rate of species $b$ at time $t$ could also depend on other species' traits.

Mathematically, the method allows the inclusion of any set of covariates that might be interesting to incorporate for evolutionary biologists; however full information on individual covariates, like traits, are rarely available – especially not on the missing species. One way to deal with this is by including an extra augmentation step and simulating full information of traits on augmented trees (Hoehna *et al.*, 2019). Another option is to use observable proxies related to e.g. trait diversity, such as different forms of phylogenetic diversity. These present interesting direction for future work.

## 2.3. MLE INFERENCE WITH MCEM USING IMPORTANCE SAMPLING

The loglikelihood of a full tree including extinct branches $x \in \mathscr{X}$ involving a total of $M$ events by extrapolating from (2.1) can easily be shown to be given by

$$\ell_x(\theta) = \sum_{i=1}^{M} \sum_{b=1}^{N_{t_i}} \left[ \log\left[\lambda_b(t_i;\theta)\mathbb{1}_{Sp}(t_i,b) + \mu_b(t_i;\theta)\mathbb{1}_{Ex}(t_i,b)\right] - \int_{t_{i-1}}^{t_i} \lambda_b(t;\theta) + \mu_b(t;\theta)\, dt \right] \tag{2.3}$$

where $\mathbb{1}_{Sp}(t_i,b) = 1$ if species $b$ speciates at time $t_i$, 0 otherwise and $\mathbb{1}_{Ex}(t_i,b) = 1$ if species $b$ becomes extinct at time $t_i$, 0 otherwise. An additional term $-\sum_{b=1}^{N_{t_M}} \int_{t_M}^{T} \lambda_b(t;\theta) + \mu_b(t;\theta)\, dt$ has to be added to the likelihood, if the final event time $t_M$ does not correspond to the present $T$. For the case when diversification rates are step-wise constant this reduces to the solutions in Wrenn (2012) and Reynolds (1973). When the full phylogenetic tree and the covariates at all times are given, we can directly maximize the loglikelihood function (2.3) to obtain the maximum likelihood estimates of the parameters (Paradis, 2005) and perform model selection to determine what factors are important for diversification. In practice, however, we almost never observe the full phylogenetic tree, but only a tree with the extant species.

### 2.3.1. DIFFICULTIES OF MLE ESTIMATION AND AN MCEM ALGORITHM

Let us denote $\mathscr{Y}$ as the space of ultrametric trees (Gavryushkin and Drummond, 2016), i.e., time-calibrated trees without extinct lineages, and $\mathscr{X}(y)$ as the space of all full trees that, when pruning all extinct species, lead to the ultrametric tree $y \in \mathscr{Y}$. Then the log likelihood of an observed, extant species only tree $y$ is given by the integral of the likelihood (2.3) over all possible full trees,

$$\ell_y(\theta) = \log \int_{\mathscr{X}(y)} \exp\left(\ell_x(\theta)\right)\, dx. \tag{2.4}$$

However, because of the complexity of the space $\mathscr{X}(y)$ a closed-form solution for equation (2.4) is not available in most cases(Gavryushkin *et al.*, 2016), making inference, or in particular, direct MLE computations difficult or impossible.

A typical method for likelihood maximization under incomplete data is the application of the expectation-maximization (EM) algorithm (Dempster *et al.*, 1977), considering the information about the extinct species as a missing data problem. In the EM algorithm, a sequence $\{\theta^{(s)}\}$ of parameter values are generated by iterating the following two steps,

**E-step** Compute the conditional expectation $Q(\theta|\theta^{(s)}) = \mathbb{E}_{\theta^{(s)}}(\ell_X(\theta)|Y = y)$,

**M-step** Choose $\theta^{(s+1)}$ to be the value of $\theta \in \Omega$ which maximizes $Q(\theta|\theta^{(s)})$.

This algorithm is run iteratively until convergence is reached. Under certain regularity conditions (Dempster *et al.*, 1977), the point of convergence can be shown to be the MLE for the incomplete data problem, i.e., maximizing $\ell_y(\theta)$.

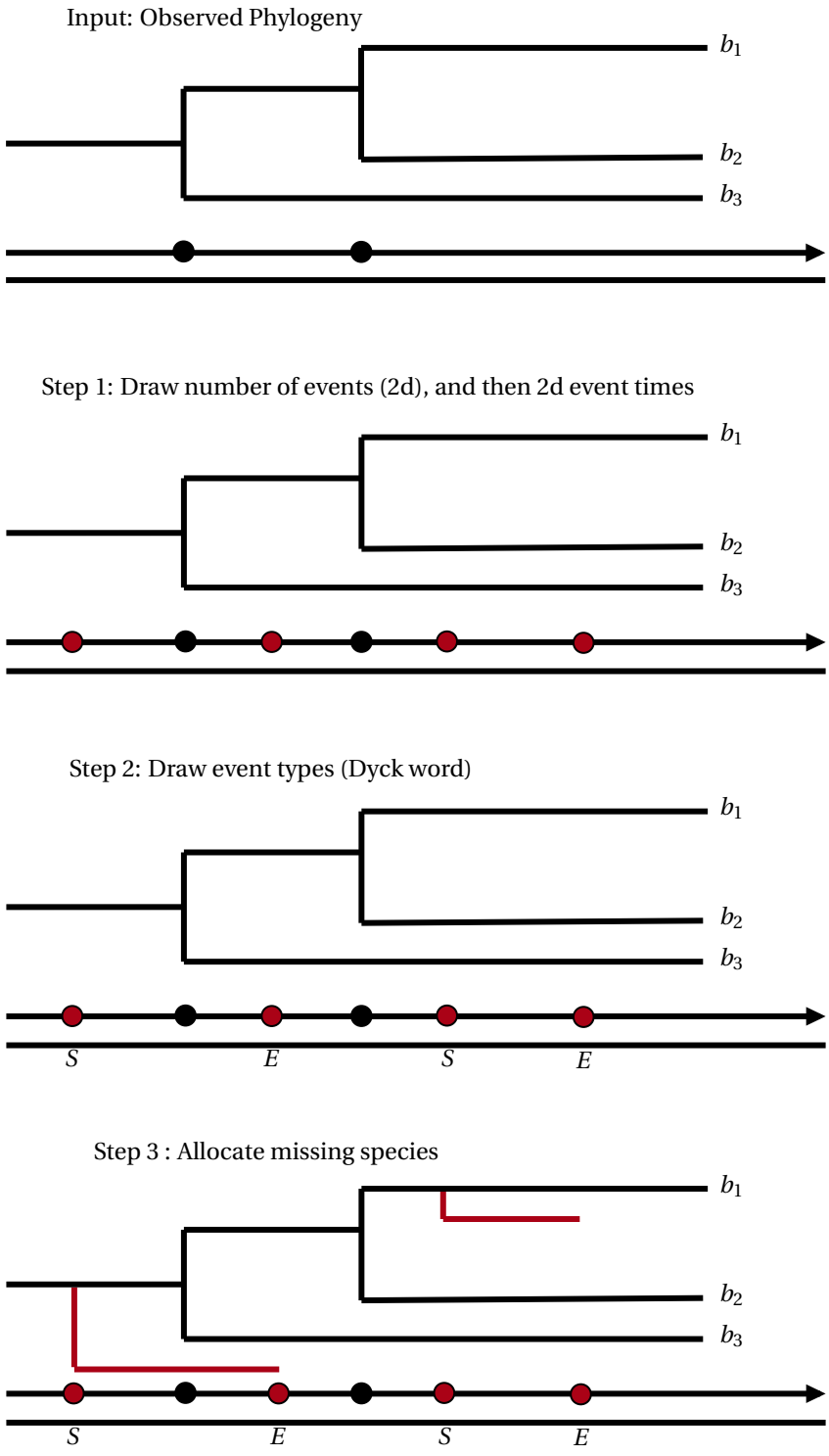



**Figure 2.2 |** The three components of our phylogenetic tree augmentation algorithm.

As in the case of equation (2.4), the calculation of $Q(\theta|\theta^*)$ does not have a closed-form due to the complexity of the space $\mathscr{X}(y)$, so approximations are needed. To perform this task we use Monte-Carlo integration (Wei and Tanner, 1990), where given a set of sampled trees $x_1, ..., x_p$ from an importance sampler distribution $g(x|y,\theta)$ we approximate $Q(\theta|\theta^*)$ by

$$\begin{aligned} Q(\theta|\theta^*) &\approx \frac{1}{p}\sum_{i=1}^{p} \ell_{x_i}(\theta)\frac{f_{X|Y}(x_i|y,\theta^*)}{g_{X|Y}(x_i|y,\theta^*)} \\ &\propto \frac{1}{p}\sum_{i=1}^{p} w_i \ell_{x_i}(\theta) \end{aligned} \tag{2.5}$$

where the importance weights are defined as $w_i = \frac{f_{X,Y}(x_i,y|\theta^*)}{g_{X|Y}(x_i|y,\theta^*)}$, using the law of conditional probabilities to obtain the proportional expression.

In the M-step we optimize (2.5) via numerical methods and the Hessian is calculated and represents the Fisher information matrix $H^{-1}$. Assuming that the errors given by the EM algorithm are independent of the Monte Carlo errors, the standard errors for the MCEM algorithm are defined as

$$SE(\hat{\theta}_i) = \sqrt{-H^{-1}_{i,i} + \frac{\text{VMCE}}{N_{EM}}} \tag{2.6}$$

where $-H^{-1}_{i,i}$ corresponds to the diagonal components of the information matrix giving the EM error, VMCE is the variance of the MC error and $N_{EM}$ is the number of MCEM iterations considered for estimation. Note that if the EM is run long enough, the second term in (2.6) goes to zero, making the information matrix the decisive value for standard errors on MCEM algorithm (McLachlan and Krishnan, 2007). With the standard errors we can construct confidence intervals for the parameters and test hypotheses about the significance of covariates of interest.

### 2.3.2. A SIMPLE IMPORTANCE SAMPLER

To sample trees we propose a tree augmentation algorithm that samples independently the three components of the tree: event types, event times and species allocations. The algorithm is shown in Figure 2.2.

***Step 1.* Generate event times and number of extinctions.** The number of extinct species $d$ and $2d$ missing event times, i.e., speciations and extinctions of these $d$ missing species are sampled uniformly in the following manner:

1. Sample the number of missing species $d$ uniformly from the discrete space $\{0, \dots, M^e\}$ where $M^e$ is a predefined ceiling, such that the probability of more than $M^e$ extinctions is extremely unlikely.

2. Sample $2d$ branching times uniformly from the continuous space $(0, T]$ and then sort them.

The probability of sampling a set of $2d$ unobserved event times $t^e = (t_1^e, \ldots, t_{2d}^e)$ for a tree of dimension $d$ is

$$g_{\text{event times}}(d, t^e) = \frac{1}{M^e + 1} \left(\frac{1}{T}\right)^{2d} (2d)!$$

Note that this scheme samples the dimension of the tree uniformly, but the size of the space of trees grows in a factorial way with the dimension of the tree. This means that the sample size required to obtain a robust Monte Carlo approximation of the integral (2.4) must be large. This is a limitation of this importance sampler, and hence it is only reliable when many extinctions are unlikely.

***Step 2. Generate event types*** We simulate a binary event chain $\tau^e = (\tau_1^e, \ldots, \tau_{2d}^e)$ assigning either S (speciation) or E (extinction) to each event time. This chain is subject to the rule that the number of Es up to any point in the chain should be less than or equal to the number of Ss in the chain up to that point. The set of allowed chains is known in the mathematical literature as the set of Dyck words and several methods for sampling Dyck words have been developed (Kasa, 2010). Furthermore, given a number of events $2d$, the number of possible Dyck words is known as the Catalan number (Zvonkin, 2014),

$$C_d = \binom{2d}{d} \frac{1}{d+1}.$$

By uniformly sampling a Dyck word $\tau^e$ of length $2d$, the probability of a specific event sequence is given by $g_{\text{events}}(\tau^e) = 1/C_d$.

***Step 3. Species allocation*** Given the missing event times and missing event types we can perform the tree allocations by sampling a parent species of each missing speciation and by defining which species, i.e., the parent species or the inserted "new species", becomes extinct at the extinction event. To sample uniformly we just need to count the number of possible trees in agreement with the event times $t^e = (t_1^e, \ldots, t_{2d}^e)$ and event types. This number, $n(\tau_{2d}^e, t_{2d}^e)$, can be calculated by starting with $n(\tau_0^e, t_0^e) = 1$ and applying the following rules when going from root to tips in the phylogenetic tree:

- For each unobserved speciation event at $t_i^e$, i.e., $\tau_i^e = S$, update $n(\tau_i^e, t_i^e)$ in the following way,

  $$n(\tau_i^e = S, t_i^e) = n(\tau_{i-1}^e, t_{i-1}^e) \times \left(2N_{t_i^-}^o + N_{t_i^-}^e\right),$$

  where $N_{t^-}^o$ is the number of observed branches just before $t$ and $N_{t^-}^e$ is the number of unobserved branches just before $t$. Note that events on observed branches count twice compared to those on unobserved branches. Intuitively, this accounts for the two eventualities following an unobserved speciation on an observed branch: either the first or the second daughter species is observed (the other one is unobserved), while for a speciation on an unobserved branch both daughter species are unobserved. A more formal argument justifying the factor of two is provided by Laudanno *et al.* (2019).

- For each unobserved extinction event at $t_i^e$, i.e., $\tau_i^e = E$, update $n(\tau_i^e, t_i^e)$ in the following way,
$$n(\tau_i^e = E, t_i^e) = n(\tau_{i-1}^e, t_{i-1}^e) \times N_{t_i^-}^e.$$

As we sample uniformly, the probability for each possible allocation $a^e$ of the $d$ missing species at the missing event times $t^e$ with Dyck word $\tau^e$ in the tree of extant species $x_{obs}$ is then given by $g_{\text{allocation}}(a^e) = \frac{1}{n(\tau_{2d}^e, t_{2d}^e)}$.

***Sampling probability of a uniformly augmented tree*** The uniform sampling probability of the augmented tree $x_{\text{unobs}} = (d, t^e, \tau^e, a^e)$ is then given by

$$g(x_{\text{unobs}}|x_{\text{obs}}, \theta) = \frac{1}{M^e + 1} \left(\frac{1}{T}\right)^{2d} (2d)! \frac{1}{C_d} \frac{1}{n(\tau_{2d}^e, t_{2d}^e)} \tag{2.7}$$

From this equation we can see how the dimension of the tree space plays an important role. For this reason, the uniform importance sampler becomes less efficient when many extinctions are likely. On the other hand, the uniform sampling scheme allows for easy implementation and quick computation, thereby making it suitable as a default sampler.

### 2.3.3. CHECKING PERFORMANCE BY COMPARING WITH DIRECT ML

To show that the MCEM works, we compared our method to the linear diversity-dependence (LDD) diversification model for which the likelihood can be calculated directly (Etienne *et al.*, 2012b). In this model speciation rates depend on diversity of the phylogenetic tree at that point. We consider the diversification model with rates

$$\lambda_b(t) = \lambda_0 - (\lambda_0 - \mu_0)\frac{N_t}{K}, \quad \mu_b(t) = \mu_0$$

where $N_t$ is the number of extant species (diversity) at time $t$ and $\theta = \{\lambda_0, \frac{\mu_0 - \lambda_0}{K}, \mu_0\}$ are model parameters. This model is a special case of our general modelling framework, defined in (2.2). We perform the MCEM routine on a clade of Malagassy birds, the so-called Vangidae clade shown in Figure 2.3, which has been analyzed in Jønsson *et al.* (2012). We replicated the routine several times with different sample sizes to observe the impact of sample size on estimation and the robustness of the method.

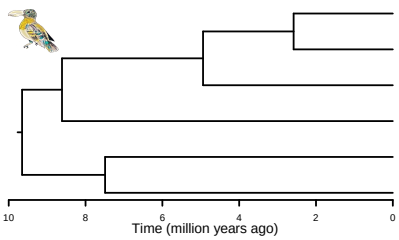



**Figure 2.3 |** Subclade of the Malagassy Vangidae, obtained from Jønsson *et al.* (2012).

In table 2.1 we show 6 replicates corresponding to 3 pairs with different sample size orders. We drop the first 1000 iterations as burn-in, and use the next 1000 MCEM iterations for parameters estimation, reporting the mean value and the standard error from equation (2.6). We observe that for small sample sizes (replicates 1 and 2) estimation is poor. For the cheapest set-up the mean effective sample size (ESS) is approximately 37 and this does not seem enough to sample in spaces with a substantial number of missing species. In this scenario, the MCEM estimates are not robust. As sample size increases we see that

inference becomes more and more accurate and matches the MLE procedure by Etienne *et al.* (2012b).

| $\overline{ESS}$ | replicate | $\hat{\theta}_1$ | $SE(\hat{\theta}_1)$ | $\hat{\theta}_2$ | $SE(\hat{\theta}_2)$ | $\hat{\theta}_3$ | $SE(\hat{\theta}_3)$ |
|---|---|---|---|---|---|---|---|
| 37 | 1 | 1.403 | 0.077 | -0.257 | 0.016 | 0.032 | 0.026 |
| 37 | 2 | 1.359 | 0.077 | -0.249 | 0.016 | 0.031 | 0.026 |
| 373 | 3 | 1.709 | 0.098 | -0.307 | 0.020 | 0.046 | 0.031 |
| 372 | 4 | 1.713 | 0.098 | -0.309 | 0.020 | 0.046 | 0.031 |
| 2970 | 5 | 1.932 | 0.127 | -0.336 | 0.026 | 0.056 | 0.033 |
| 2987 | 6 | 1.892 | 0.121 | -0.328 | 0.025 | 0.056 | 0.033 |
| MLE | | 1.937 | | -0.326 | | 0.060 | |

**Table 2.1 |** MCEM estimation for 3 different samples sizes, with 2 replicates each. The first column is the mean of the effective sample size over the 1000 iterations considered. The last row is the MLE directly calculated by computing the likelihood (Etienne *et al.*, 2012b). Estimated values are for the linear DD model with $\theta_1 = \lambda_0$, $\theta_2 = (\mu_0 - \lambda_0)/K$ and $\theta_3 = \mu_0$.

These replicates are also summarized in Figure 2.4 where we show a visualization of the dynamical MCEM parameter estimation for $\log \lambda_0$ corresponding to the logarithm of the initial speciation rate at stem age. The dashed black line indicates the true MLE. We see in all 6 cases that estimations go quickly to the true MLE with a stable behaviour after a couple of hundred iterations. To visually compare biases and variation through different sample sizes we show the replicates for small sample sizes until the 2000th and 2500th MCEM iteration. We clearly see that for higher sample sizes bias and variation decrease.

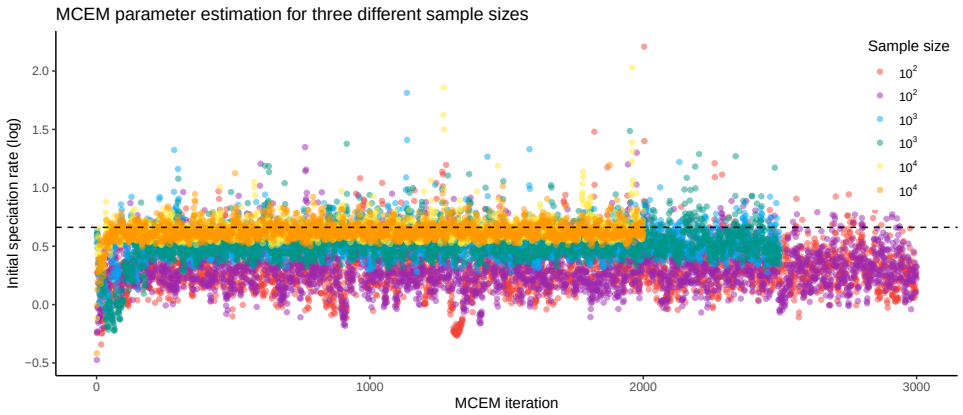



**Figure 2.4 |** MCEM applied to the tree of Figure 3 under the LDD diversification model. Evolution of the estimate of the first parameter, the initial speciation rate $\theta_1 = \lambda_0$ through EM iterations. We plot 6 replicates: 2 for 3 different sample sizes. For better visualization we cut higher sample sizes at iteration 2000 and 2500.

Note that the effective sample size is between 30% and 40% in these cases. An efficient importance sampler with 100% effective sample size is a priority for future publications in order to apply the method to larger phylogenetic trees.

## 2.4. DIVERSITY-DEPENDENCE: DIVERSITY OR PHYLODIVERSITY?

Phylodiversity is defined as the total branch length of extant species of a tree, and it has been proposed as an alternative to diversity in conservation ecology (Faith, 1992). Figure 2.5 shows phylodiversity and diversity through time for a simple example tree.

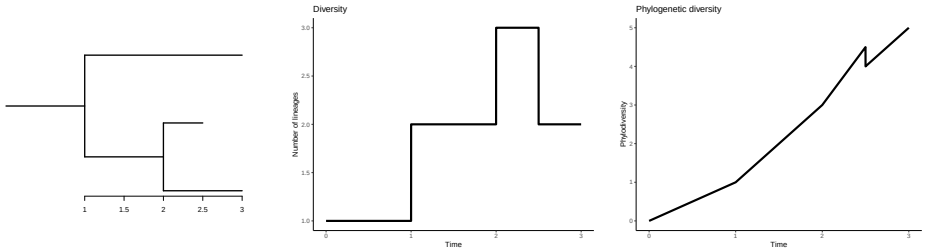



**Figure 2.5 |** Example of a simple tree with one extinction. The two panels on the right show the difference between diversity and phylodiversity through time.

As an illustration of the flexibility of our method we now consider a model similar to diversity-dependence introduced in the previous section, but with dependence on phylodiversity $P_t$ instead of $N_t$. Diversity-dependence has been detected in a Vangidae clade (Jønsson *et al.*, 2012) and we would like to extend the analysis to check if phylodiversity-dependence (LPD) is a more suitable factor in diversification of Vangidae than diversity-dependence (LDD). In addition to these two models, which both assume linear dependence of speciation rate on diversity or phylodiversity, we consider the exponential diversity dependence (EDD) and exponential phylogenetic diversity (EPD) models. The exponential models use the log-link function common in the statistical literature, rather than the identity link suggested by the evolutionary biology literature. Table 2.2 shows the parameter definitions for the four models tested on the phylogenetic tree of the Vangidae.

| Model | $\lambda_b(t)$ | $\theta_1$ | $\theta_2$ | $\theta_3$ |
|---|---|---|---|---|
| LDD | $\lambda_0 - (\lambda_0 - \mu_0)\frac{N_t}{K}$ | $\lambda_0$ | $-(\lambda_0 - \mu_0)\frac{1}{K}$ | $\mu_0$ |
| LPD | $\lambda_0 - (\lambda_0 - \mu_0)\frac{P_t}{K}$ | $\lambda_0$ | $-(\lambda_0 - \mu_0)\frac{1}{K}$ | $\mu_0$ |
| EDD | $\lambda_0 e^{-aN_t}$ | $ln(\lambda_0)$ | $-a$ | $\mu_0$ |
| EPD | $\lambda_0 e^{-aP_t}$ | $ln(\lambda_0)$ | $-a$ | $\mu_0$ |

**Table 2.2 |** Four diversity-dependent diversification models, where speciation rate depends on diversity or phylodiversity, either linearly or exponentially. All models assume constant extinction rate and have 3 parameters to be estimated.

We performed the MCEM routine for each of the four diversification models, obtaining the ML estimates of the parameters and calculating Monte-Carlo estimation for the likelihood function and the corresponding AIC values (Wit *et al.*, 2012). Interestingly, we found that phylodiversity models do not performs better than ordinary diversity models, but there is an improvement of the exponential diversity-dependence model over the

linear DD model. Table 2.3 shows the inference results for each of the four diversification models.

To get an idea of the computational cost of the method we include, next to table 2.3, a plot of computing times (for one MCEM iteration) as a function of Monte Carlo sample size for PD and DD models starting at their respective MLE values reported in the table. The values are average of 100 replicates performed in an ordinary computer. From the plot we can see that for our example tree each iteration takes a couple of minutes for large Monte Carlo sample size, which means that the whole routine should take few hours at most. We also see that the computing times increases linearly with the MC sample size.

| Model | $\theta_1$ | $\theta_2$ | $\theta_3$ | loglikelihood | AIC |
|---|---|---|---|---|---|
| LDD | 1.94 | -0.33 | 0.06 | -11.36 | 28.72 |
| LPD | 0.31 | -0.01 | 0.04 | -14.37 | 34.74 |
| **EDD** | **2.58** | **-1.02** | **0.04** | **-11.19** | **28.37** |
| EPD | -0.28 | -0.04 | 0.13 | -13.44 | 32.89 |

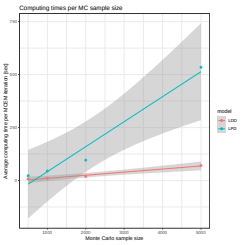



**Table 2.3 |** Parameter estimation of the four diversity-dependent models of Table 2.2 when applied to the Vanga tree of Figure 2.3, including Monte-Carlo approximations of the loglikelihood and AIC. Next to the table we see the plot of average computing times per MCEM iteration (in seconds) for DD and PD models at their respective MLE.

We conclude that the best model in this analysis is an EDD model with parameters $\theta_1 = 2.58(0.96)$, $\theta_2 = -1.02(0.25)$, $\theta_3 = 0.04(0.03)$, suggesting an exponential decreasing speciation rate with a exponential decay constant close to 1, given by $\theta_2$. We found an initial speciation rate of approximately 4.85 species per million years which decreases until 0.03 at the present time. This indeed suggests that the diversification process of this Vangidae clade in Madagascar has slowed down dramatically over the past 10 million years. Moreover, the extinction rate of 0.04 species per million years suggests that the clade has now reached a stable diversification behaviour, whereby any further speciations will tend to be offset by extinctions.

## 2.5. DISCUSSION

We have presented a flexible method for testing a broad variety of diversification models in phylogenetic analysis and provided some simple examples. This is a first step towards a robust general methodology to identify potential factors in diversification processes from phylogenetic trees.

The unobserved extinct species turn the inference problem naturally into a problem that can be approached by means of an EM algorithm. Given the complexity of the E-step, a Monte Carlo importance sampler has been proposed, involving a uniform importance sampler. Given the computational simplicity both in terms of sampling and calculation of uniform samplers this may be a convenient option for small sized trees, where more sophisticated importance samplers, involving the underlying non-homogenous Poisson processes, would not necessarily improve efficiency. As in the case of Vangidae clade

where few missing species are likely, we found that the uniform importance sampler leads to accurate estimation. However, the performance of our uniform importance sampler deteriorates as the dimension of the phylogenetic tree increases. In order to apply this method on high-dimensional trees a more efficient importance sampler should be carefully chosen. This we will leave for future work.

Current approaches perform inference by means of likelihood maximization, which requires that formulas for the likelihood must be derived on a case-by-case basis. Here, we consider a general class of models that include an augmentation step inside an EM algorithm, thereby avoiding direct likelihood calculation and thus allowing inference for a wide variety of diversification models.

In principle, in cases when full information of covariates is still missing after the augmentation step, extensions of the augmentation procedure are possible. However, this is beyond the scope of the current paper.

Moreover, to increase efficiency alternatives to MCEM algorithms may be considered, such as the stochastic approximation version of the EM algorithm (SAEM) (Delyon *et al.*, 1999) or a Bayesian approach (Richardson and Green, 1997). In both cases the algorithm could make use of the previous MC samples, thereby improving efficiency at some computational cost.

Even though in this paper we only refer to the context of a diversification process of ecological species, a phylogenetic tree is used in many other fields to describe other kinds of processes, such as language evolution (Greenhill *et al.*, 2010) and cultural diversification (Mace and Holden, 2005). Therefore, the method that we have developed in this paper is potentially useful for inferring the underlying driving process of such branching processes.

# 3

# DETECTING PHYLODIVERSITY-DEPENDENT DIVERSIFICATION WITH A NOVEL PHYLOGENETIC INFERENCE FRAMEWORK

*In general, however, constructing data augmentation schemes that result in both simple and fast algorithms is a matter of art in that successful strategies vary greatly with the models being considered.*

David A van Dyk and Xiao-Li Meng

Richter F, Wit EC, Etienne RS. In preparation.

## ABSTRACT

*Diversity-dependent diversification models have been extensively used during the last decade in phylogenetic analysis to study the effect of ecological limits and feedback of community structure on species diversification processes, such as speciation and extinction. Current diversity-dependent diversification models characterise ecological limits by carrying capacities for species richness. Such ecological limits have been justified by niche filling arguments: as species diversity increases, the number of available niches for diversification decreases.*

*However, as species diversify they may diverge from one another phenotypically, which may open new niches for new species. Alternatively, this phenotypic divergence may not affect the species diversification process or even inhibit further diversification. Hence, it seems natural to explore the consequences of phylogenetic diversity-dependent (or phylodiversity-dependent) diversification. Current likelihood methods for estimating diversity-dependent diversification parameters cannot be used for this, as phylodiversity is continuously changing as time progresses and species form and become extinct.*

*In this chapter, we present a new method based on Monte Carlo Expectation-Maximization (MCEM), designed to perform statistical inference on a general class of species diversification models and implemented in the R package* `emphasis`. *We use the method to fit phylodiversity-dependent diversification models to 14 phylogenies, and compare the results to the fit of a richness-dependent diversification model. We find that in a number of phylogenies, phylogenetic divergence indeed spurs speciation even though species richness reduces it. Not only do we thus shine new light on diversity-dependent diversification, we also argue that our inference framework can handle a large class of diversification models for which currently no inference method exists.*

## 3.1. INTRODUCTION

The hypothesis of diversity-dependent diversification posits that diversification processes at macro-evolutionary scales are affected by community structure, and particularly by diversity (Gould *et al.*, 1977; Walker and Valentine, 1984). One of the underlying ideas is that there are ecological limits to diversity (there is a limited number of niches that can be filled with species) and hence to diversification (Rabosky, 2009). The hypothesis has been extensively studied both empirically and theoretically (Rabosky and Hurlbert, 2015; Etienne *et al.*, 2016; Morlon, 2014; Jønsson *et al.*, 2012; Condamine, 2018; Gibb *et al.*, 2016; Cunha *et al.*, 2017; Pouchon *et al.*, 2018; Chen *et al.*, 2017; Pinto-Ledezma *et al.*, 2017; McGuire *et al.*, 2014; Pyron and Wiens, 2013; Xu and Etienne, 2018; Liow *et al.*, 2010; Herrera-Alsina *et al.*, 2018). However, currently developed inference models for detecting diversity-dependent diversification from molecular phylogenies consider only species richness as a proxy for diversity (Etienne *et al.*, 2012b).

Phylogenetic diversity, quantifying the genetic differences among a group of species, has been identified as a key feature of diversity (Kling *et al.*, 2018; Scheiner *et al.*, 2017) to be taken into account in conservation biology (Laity *et al.*, 2015; Faith and Baker, 2006) (but see Cantalapiedra *et al.* (2019); Mazel *et al.* (2018)), community ecology (Stadler *et al.*, 2017; Tucker *et al.*, 2016; Webb *et al.*, 2006; Violle *et al.*, 2011), evolutionary biology (Kling *et al.*, 2018) and the intersection of these fields. Phylogenetic diversity, or phylodiversity, provides a different perspective on diversity and ecological limits. Whereas species richness models suggest that as species diverge there may be less opportunity to speciate further as the growing phenotypic space between species leaves less room to be occupied, one may argue, however, that the divergence provides access to more space to speciate into. Hence, extending diversity-dependence to phylodiversity and developing methods to infer such phylodiversity-dependent diversification from molecular phylogenies seems worthwhile. It will allow us to consider the dynamic nature of ecological limits (Costa *et al.*, 2008; Lister, 1976; Soininen *et al.*, 2011) and thus relax the assumption of fixed limits (Etienne *et al.*, 2012b; Marshall and Quental, 2016). In this chapter we develop such an extension.

The incorporation of phylogenetic diversity is not possible with the current simulation-free methods for inferring diversity-dependent diversification using the Q-approach introduced by Etienne *et al.* (2012b) and Laudanno *et al.* (2020b) and implemented in the R package `DDD`. This method is based on a hidden Markov model approach whereby the probability of an extant-species tree is integrated over the infinite set of complete trees compatible with it, i.e., the trees that also contain now-extinct species. The method relies on the assumption that only the number of species at any point in time affects the diversification rates, and therefore does not depend on tree topology. Phylogenetic diversity, defined as the sum of the lengths of all branches in a phylogenetic tree (Faith, 1992), highly depends on the topology of the tree as well as the branching times. Hence, a new methodology is needed to incorporate topological characteristics of the diversification processes.

To do so we generalize a recently developed statistical framework (Richter *et al.*, 2020) based on Monte Carlo Expectation-Maximization (MCEM) that allows inference on a general class of diversification models, including models with phylodiversity-dependent

diversification. In this EMPHASIS (Expectation-Maximization in PHylogenetic Analysis with Simulations and Importance Sampling) framework, maximum likelihood estimation is performed on an augmented data set generated by Monte Carlo simulations.

This general class of Species Diversification Models (SDM) contains a broad spectrum of scenarios considered in the literature where rates can be constant (Nee *et al.*, 1994), be related to the age of the species (Hagen *et al.*, 2015), to the (changing) paleo-enviroment (Descombes *et al.*, 2018), to geographic patterns (Goldberg *et al.*, 2011) or to temperature and diversity (Condamine *et al.*, 2019), just to name a few. For each of these models specific likelihood formulas have been derived (implemented in different packages), but our new method can handle them in a single framework, and also applies to combinations of these models for which no such likelihood formula is available and is often impossible to derive or compute numerically. It also applies to new models such as the phylodiversity-dependent models discussed in detail here, and other models with possibly complex interactions between ecological factors and macroevolution, thereby opening endless opportunities for macroevolutionary diversification analysis. The main challenge of our framework is computationally: the Monte Carlo integration is very demanding. In this chapter we therefore provide a method to perform this integration efficiently.

We illustrate our inference method by applying it to 14 phylogenies, comparing a phylodiversity-dependent diversification model to a diversity-dependent diversification model. We generally find little difference between these two models, although the phylodiversity-dependent diversification model provides an additional narrative for the evolution of global speciation through time in several cases.

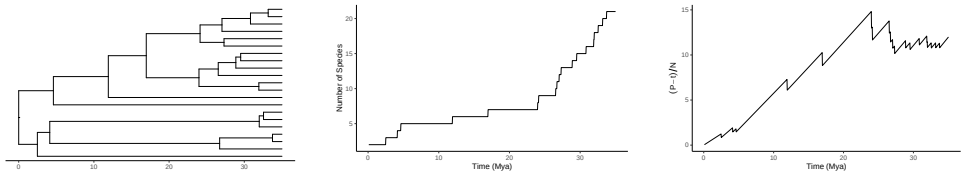
## 3.2. DIVERSITY-DEPENDENT DIVERSIFICATION MODELS

Diversity-dependent species diversification models are typically used to quantify the effect that diversity has on diversification (Condamine *et al.*, 2019; Etienne *et al.*, 2012a; Cunha *et al.*, 2017; Etienne and Haegeman, 2012; Foote *et al.*, 2018). Under the classical linear diversity-dependent diversification (LDD) model, it is assumed that speciation rate is a linear function of species richness:

$$\lambda_t = \lambda_0 + \beta_N N_t; \qquad \mu_t = \mu_0, \tag{3.1}$$

where $\lambda_t$ is the per species speciation rate, $N_t$ the number of species and $\mu_t$ represents the per species extinction rate, at time $t$. Assuming that $\lambda_0$ is positive, if $\beta_N$ is negative the quantity $K' = -\lambda_0\beta_N$ is called the carrying capacity, which denotes the value for which a clade approaches a niche limit and consequently experiences a slow-down in speciation. If $\beta_N = 0$, then the model reduces to the diversity-independent diversification model, i.e. the constant-rate model.

Phylogenetic diversity is recognised as a critical feature of diversity to take into consideration in several fields such as conservation ecology, macroecology and macroevolution. However, so far, it has been studied mostly in a qualitative way, and only as a single number at the present instead of considering it as a dynamical quantity that changes through macroevolutionary time. Current diversity-dependent diversification models do not consider phylodiversity, and assume that diversity slows down diversification (e.g. due




**(a)** Phylogeny of Megapodidae **(b)** Number of species. **(c)** Phylogenetic diversity per species

**Figure 3.1 |** Phylogeny, number of species and phylogenetic diversity per species.

to niche filling), while qualitative studies suggest that diversity can spur diversification (Jarne *et al.*, 2017; Hamilton *et al.*, 2020). Likelihood-based inference approaches ignore phylodiversity; they fully describe the processes by considering the probability that the clade has $N_t$ lineages at time $t$, but ignore the topology of the trees. We here introduce a generalised diversity-dependent diversification model, i.e., a phylodiversity-dependent diversification (LPD) model, where we assume that the speciation rate also depends on the phylogenetic diversity per species:

$$\lambda_{t;\beta} = \lambda_0 + \beta_N N_t + \beta_P \frac{P_t - t}{N_t}; \qquad \mu_{t;\beta} = \beta_0 \tag{3.2}$$

where $P_t$ is the phylogenetic diversity at time $t$ defined as the total branch length. The quantity $\frac{P_t - t}{N_t}$ corresponds to the phylogenetic diversity per species at time $t$. Note that by subtracting $t$ from the phylogenetic diversity $P_t$, the phylogenetic diversity per species remains 0 for a single species. In Figure 3.1, an example tree is plotted, with the species richness though time and the phylogenetic diversity per species through time.

In this chapter, we make use of the statistical methods described in chapter two, combined with an efficient importance sampler, in order to perform statistical inference assuming diversification dynamics given by the LPD model and compare it with the diversification dynamics given by the simple LDD model. In this way, we quantify the signal that phylodiversity leaves in species diversification and evaluate if its incorporation in diversity-dependent diversification models is promising for further studies.

## 3.3. MATERIALS AND METHODS

Phylogenetic trees are branching diagrams, reconstructed from DNA sequences, representing the evolutionary history of species diversification (Kapli *et al.*, 2020). Mathematically, they are represented by a discrete part given by the topology of the tree and a continuous part given by its branching times. We define a tree $x = \{\mathbf{t}, \tau\}$ as a combination of branching times and topology. More precisely, the branching times are defined by a chronological sequence vector of times $\mathbf{t} = \{t_0, t_1, t_2, ..., t_p\}$, with $t_0 = 0$ and $t_p$ being the present time. The topology is defined by a succession of allocation values which can be characterized in multiple ways such as in network or matrix notation. Here, we consider the succession of species names $s_1^*, s_2^*, ..., s_{p-1}^*$ to be the species that diversified (or became extinct) at branching time $t_i$. Moreover, we define the subsets of subindex $\mathscr{C}_x \subset \{1, ..., p-1\}$ and $\mathscr{E}_x \subset \{1, ..., p-1\}$ to be the indices corresponding to speciation and

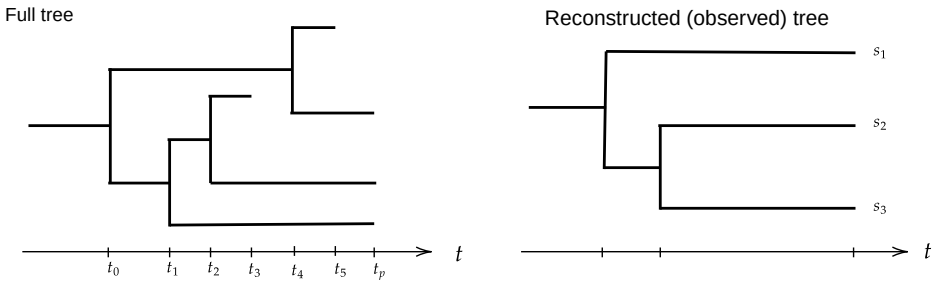



**Figure 3.2 |** Full phylogenetic trees (left) and the corresponding reconstructed tree (right). Each branch represents a species.

extinction events, respectively. This means that if $i \in \mathscr{C}_x$ then $t_i$ is a branching time corresponding to a speciation event while if $i \in \mathscr{E}_x$ then $t_i$ is a branching time corresponding to an extinction event.

Figure 3.2 shows a representation of a tree describing a full evolutionary process (speciation and extinction events), and the corresponding reconstructed tree, considering the evolutionary history of currently extant species. In ths case $\mathscr{C}_x = \{0,1,2,4\}$ and $\mathscr{E}_x = \{3,5\}$.

Throughout this chapter we consider the extant species trees to be accurate (i.e., no uncertainty in branching times or topology). Statistically, extant species trees are our observed data and extinct species are usually latent or unobserved variables, which in the case of diversity-dependent diversification also affect diversification rates.

### 3.3.1. DIVERSIFICATION OF SPECIES AS A POINT PROCESS

We consider the species diversification process as a general Point Process where each species has a waiting time to speciate into two daughter species that follows an exponential probability distribution with rate $\lambda_{t,s|\theta}$, for any time $t$, species $s$ and parameters $\theta$. Species can also become extinct with an exponential distribution with rate $\mu_{t,s|\theta}$ for the waiting time to extinction. We will denote the set of extant species at time $t$ by $\mathscr{S}_t = \{s_1, ..., s_{N_t}\}$, and the number of extant species at time $t$ by $N_t$. These quantities are described by a Non-Homogenous Poisson Process (NHPP) (Daley and Vere-Jones, 2007). Typically, we consider $\lambda_{t,s|\theta} = g\left(\sum_i \theta_i v_{i,t,s}\right)$ for a set of covariates $v_{i,t,s}$ and a link function $g : \mathbb{R} \to \mathbb{R}$ (Dobson and Barnett, 2008). The loglikelihood function of the full process represented by a complete tree (Figure 3.2, left) is given by

$$\ell_x(\theta) = \sum_{\mathscr{C}_x} \log\left(\lambda_{t_i,s_i^*|\theta}\right) + \sum_{\mathscr{E}_x} \log\left(\mu_{t_i,s_i^*|\theta}\right) - \sum_{i=1}^{p} \left[ \int_{t_{i-1}}^{t_i} \sum_{s \in \mathscr{S}_{t_i}} (\lambda_{t,s|\theta} + \mu_{t,s|\theta})\, dt \right] \tag{3.3}$$

Phylogenies that are derived from molecular data (e.g. DNA sequences) are, however, not full trees, as they do not contain the extinct species (Figure 3.2 right). The likelihood for an observed tree can be written in terms of the likelihood of compatible full trees. In principle, this is simply the integration over all possible full trees that are in agreement with the observed tree $x_{obs}$:

$$f(x_{obs}|\theta) = \int_{x \in \mathscr{X}(x_{obs})} \exp(\ell_x(\theta|x_{obs}))dx \tag{3.4}$$

This integration is usually impossible to compute in practice for most diversification models. Here, we present a method where maximum likelihood estimation is possible without calculating directly the likelihood function (3.4), by implementing a combination of statistical inference and a data augmentation algorithm.

### 3.3.2. THE EMPHASIS STATISTICAL FRAMEWORK

Our statistical framework is a generalisation of that of Richter *et al.* (2020), which makes use of an Expectation-Maximization algorithm for maximising the likelihood (Dempster *et al.*, 1977). The EM algorithm is an iterative procedure consisting of two steps: the E-step and the M-step. Starting from an initial value for the parameters, the E-step involves computing the expected loglikelihood of the observed tree for the given parameters and the M-step involves computing the parameters that maximise that expectation of the loglikelihood. Each iteration the parameters are updated with the values obtained in the M-step of the previous iteration. The E- and M-steps are run iteratively until convergence is reached. The parameters thus obtained have been shown to be the maximum likelihood estimators (Dempster *et al.*, 1977).

Because the expectation in the E-step cannot be computed exactly (or numerically) due to the high dimensionality of the space of complete trees, Richter *et al.* (2020) proposed to use a stochastic approximation and data augmentation (Tanner and Wong, 1987), specifically a Monte-Carlo method (Chan and Ledolter, 1995) in combination with importance sampling (Glynn and Iglehart, 1989) in the E-step of their EM-algorithm (McLachlan and Krishnan, 2007), and calculated

$$\begin{aligned} Q_\theta = \mathbb{E}_{\theta^*}\left[\ell_x(\theta) \mid x_{obs}\right] &\approx \frac{1}{N} \sum_{x_i \sim f(x_i|\theta, x_{obs})} \ell_{x_i}(\theta) \\ &= \frac{1}{N} \sum_{x_i \sim f_\alpha(x_i|\theta, x_{obs})} \ell_{x_i}(\theta)\, w_i \end{aligned} \tag{3.5}$$

where

$$w_i = \frac{f(x_i|\theta, x_{obs})}{f_\alpha(x_i|\theta, x_{obs})} \tag{3.6}$$

are called the *importance weights* and are highly dependent of the importance sampler $f_\alpha$. The importance weights reflects how accurate the data augmentation is in comparison with the desired distribution, importance weights equal to 1 shows that the importance sampler is the same distribution as the likelihood $f$, distribution that generates the process of interest. The data augmentation scheme used by Richter *et al.* (2020) was

mathematically correct but computationally inefficient, as the paper was aimed at the conceptual framework rather than performance. Here we present an improved version of the framework, hereafter called *emphasis*, with a very efficient data augmentation scheme, because the choice of data augmentation scheme is crucial for computational performance (Van Dyk and Meng, 2001).

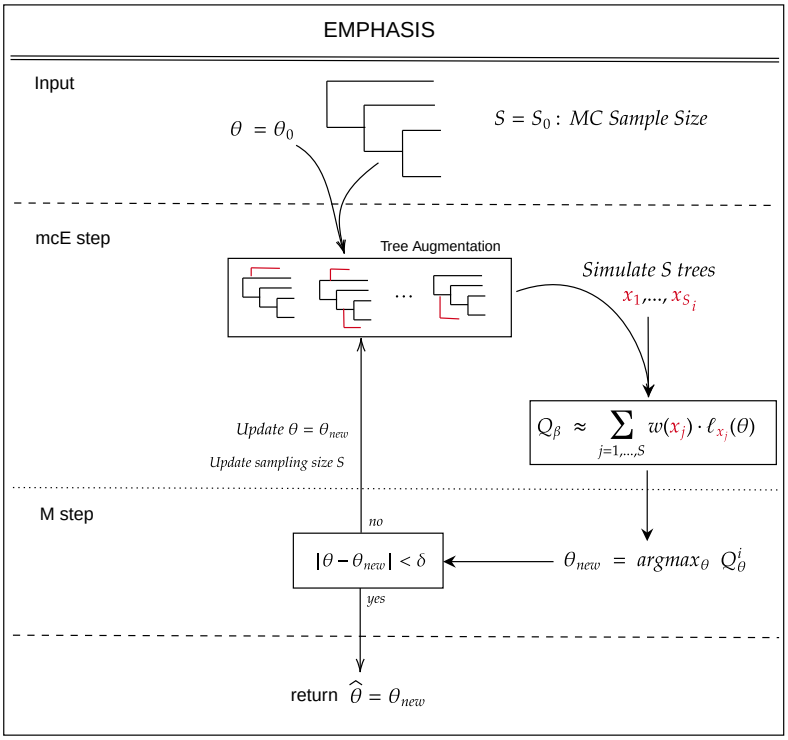



**Figure 3.3 |** Monte-Carlo EM algorithm diagram in the context of phylogenetic trees.

### 3.3.3. AUGMENTATION OF OBSERVED TREES, A NOVEL IMPORTANCE SAMPLER FOR PHYLOGENETIC INFERENCE

Richter *et al.* (2020) presents an MCEM algorithm where trees are augmented by drawing uniformly the number of branching events and its corresponding branching times. The method worked well for small trees, but the variance of the estimates grows fast as the tree gets larger for constant number of samples, making the method computationally intractable for medium-sized to large clades. This is due to the curse of dimensionality, i.e., the problem of exploring high-dimensional spaces efficiently (Friedman, 1997). Our proposed alternative for the data augmentation algorithm augments trees according to the underlying diversification model, encouraging samples in the regions of parameter space that are likely under the proposed SDM.

To sample the extinct species in the tree, we approximate the diversification process of extinct lineages conditional on the extant species in the data as a birth-death process with rates

$$\lambda^m_{t,s|\theta} = \lambda_{t,s|\theta} P_\alpha(t, t_p), \quad \mu^m_{t,s|\theta} = \frac{\mu_{t,s|\theta}}{P_\alpha(t, t_p)} \tag{3.7}$$

where $P_\alpha(t, t_p)$ is an approximation of the probability that a species observed at time $t$ will not have any descendants at time $t_p$. Kendall (1948) showed that, for lineage-independent models, the exact probability is given by

$$P_0(t_c, t_p) = \frac{\int_{t_c}^{t_p} \mu_{\tau,s|\theta} e^{-\int_{t_c}^{\tau} (\lambda_{r,s|\theta} - \mu_{r,s|\theta}) dr} d\tau}{1 + \int_{t_c}^{t_p} \mu_{\tau,s|\theta} e^{-\int_{t_c}^{\tau} (\lambda_{r,s|\theta} - \mu_{r,s|\theta}) ds} d\tau} \tag{3.8}$$

Note that the probability depends on information on $\lambda_{ts|\theta}$ and $\mu_{ts|\theta}$ for $t_c < t < t_p$. For constant rates this information is available and calculation of Eq. 3.8 is easy. However, for most SDM information on the full process is not available. For instance, in the case of diversity-dependent diversification models the quantity $N_t$ is unknown.

We augment the observed tree with hidden speciation events. These events can be allocated to all lineages, but not with equal probability. Speciation events occurring on an observed lineage have twice the weight of speciation events occurring on an unobserved lineage (Etienne *et al.*, 2012b). Figure 3.4 shows an example when a tree is augmented with a new speciation event at a time that there are two extant lineages and one extinct lineage. In that case, there are five possible allocations. More generally, there are $N^e_{t^-} + 2N^o_{t^-}$ possible allocations, where $N^e_{t^-}$ is the number of currently extinct lineages alive just before time $t$ and $N^o_{t^-}$ is the number of currently extant lineages just before time $t$. Therefore, we can compute the probability distribution for the waiting times for the augmented speciation events (which we will call missing speciation events) considering the $N^e_{t^-} + 2N^o_{t^-}$ non-homogenous Poisson processes together. Because the minimum waiting time for exponential distributed processes is also an exponential process, given a time $t_0$, the waiting time for the first missing speciation event to occur is given by an exponential distribution with rate

$$\sigma_{t|\theta} = \sum_{s \in \mathscr{S}^m_t} \lambda^m_{t,s|\theta} + 2 \sum_{s \in \mathscr{S}^0_t} \lambda^m_{t,s|\theta}$$

where $\mathscr{S}^m_t$ and $\mathscr{S}^o_t$ are the sets of observed and missing species at time $t$ respectively. Hence, the probability density of the waiting time for any speciation to occur at time $t$, starting the process at initial time $t_i$, is a non-homogeneous exponential distribution with rate $\sigma_{t|\theta}$, that is

$$f_B(t_c | \, t_i \, , \theta) = \sigma_{t|\theta} e^{-\int_{t_i}^{t_c} \sigma_{t|\theta} dt}.$$

Once a missing speciation event has occurred, the new lineage needs to get an allocation and a extinction time assigned to be included in the tree. In a model where speciation

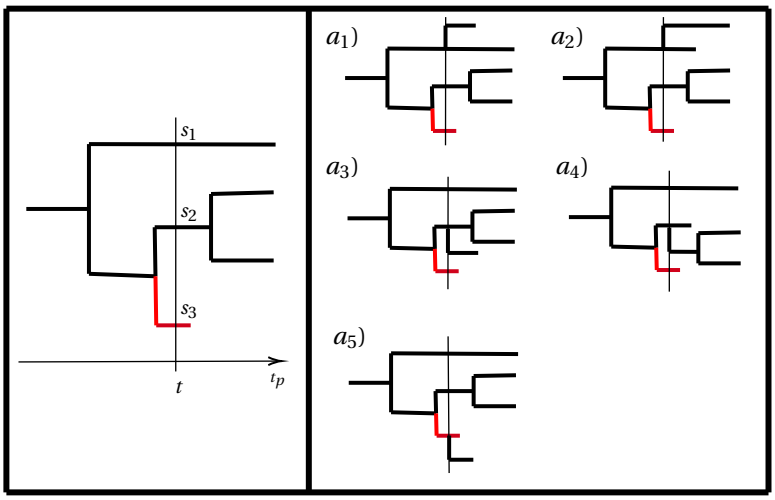



**Figure 3.4 |** Phylogenetic tree with 2 observed species and 1 missing species at time $t$. When a new species is created there are $2N_0 + N_m$ possible allocations, in this case $2 * 2 + 1$.

rates are the same for all lineages, all allocations have the same probability

$$\mathbb{P}_A(\tau|t_c,\theta) = \frac{1}{N^e_{t^-} + 2N^o_{t^-}}.$$

The lineage produced at the missing speciation event must become extinct before the present. The extinction time of the species $s$ born at time $t_c$ is a random variable with a density distribution that is conditioned on extinction occurring before time $t_p$,

$$f(t_e|s,t_c) = \mu_{t_e,s|\theta} \frac{e^{-\int_{t_c}^{t_e} \mu_{q,s|\theta} dq}}{1 - e^{-\int_{t_c}^{t_p} \mu_{q,s|\theta} dq}}.$$

Note that this probability also depends on the extinction rate of the full process (i.e., at times later than $t_c$), which is not always available, as it may depend, for example, on diversity at those later times. Hence, we propose to sample the extinction time from the truncated distribution

$$f_D(t_e|s,t,\theta) = \mu_{t,s|\theta} \frac{e^{-\mu_{t,s|\theta}(t_e - t)}}{1 - e^{-\mu_{t,s|\theta}(t_p - t)}}.$$

The full sampling probability of the missing part of a tree under this scheme is then given as

$$f_m(x|\theta) = \prod_{i \in \mathcal{M}_\tau} f_B(t_i|\ t_{i-1}) \mathbb{P}_A(a_i|t_i) f_D(t^e_i|a_i,t_i). \tag{3.9}$$

THE DATA AUGMENTATION ALGORITHM (DAA)

The main idea for our proposed data augmentation algorithm is to replace $P_\alpha(t)$ by the probability that the newly created species (and not the entire clade that will descend from it) will become extinct before the present time

$$P_1(t_c, t_p) \;=\; 1 - e^{-\mu_{t_c}(t_p - t_c)}$$

Thus, we consider the evolutionary process with diversification rates

$$\lambda^m_{t,s|\theta} \;= \lambda_{t,s|\theta}(1 - e^{-\mu_{t_c}(t_p - t_c)}), \quad \mu^m_{t,s|\theta} \;= \frac{\mu_{t,s|\theta}}{(1 - e^{-\mu_{t_c}(t_p - t_c)})} \tag{3.10}$$

The algorithm is based on a Gillespie-type simulation algorithm which is computationally simple and feasible relatively simple digital computer algorithm (Gillespie, 1976; Kieu, 2018).

The algorithm proceeds as follows:

1. Input: Set $t_0 = 0$, $i \;=\; 1$.
2. Draw a **missing speciation time** $t$ from distribution

$$f_B(t|\; t_i\;, \theta) \;= \sigma_{\,t|\theta}\, e^{\;-\int_{t_i}^{t} \sigma_{\,t|\theta}\, dt}.$$

   where

$$\sigma_{\,t|\theta} = \sum_{s \in \mathscr{S}_t^m} \lambda_{t,s|\theta}(1 - e^{-\mu_t(t_p - t)}) + 2 \sum_{s \in \mathscr{S}_t^0} \lambda_{t,s|\theta}(1 - e^{-\mu_t(t_p - t)})$$

3. Draw an **allocation** for the species from distribution

$$P_A(\tau|t, \theta) \;= \frac{1}{N^e_{t^-} + 2N^o_{t^-}}$$

4. Draw the corresponding **extinction time** from distribution

$$f_D(t_e|s(\tau), t, \theta) = \;\mu_{t,s|\theta} \frac{e^{-\mu_{t,s|\theta}(t_e - t)}}{1 - e^{-\mu_{t,s|\theta}(t_p - t)}}$$

5. Set $t_i \;=\; t$, if $t_i < t_p$ update the tree with the new species (speciation time, extinction time and allocation) and go to step 2; if $t_i > t_p$ stop the algorithm and return the augmented tree.

An interpretation of this process is as follows:

- We observe a process with varying extinction rates, but as soon as a new missing species arises the extinction rate of that species is fixed throughout the rest of the process.
- When allocating a new species we assume that all possible allocations have a uniform probability distribution.

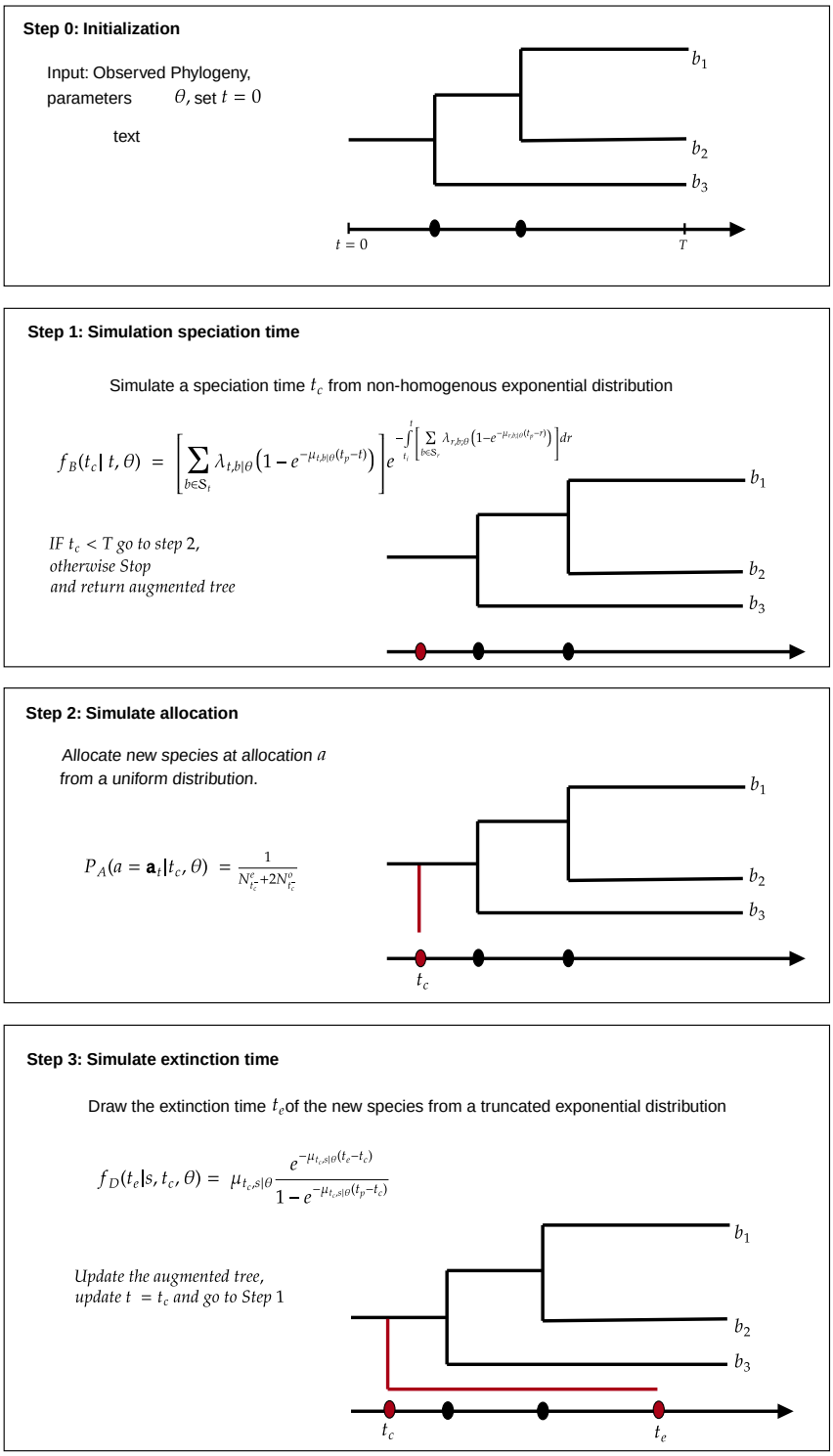



**Figure 3.5 |** Tree augmentation algorithm based on the underlying non-homogeneous Poisson process.

- We consider alternative probabilities $P_\alpha(t,T)$, thus indexed by $\alpha$, instead of the probability of not having any descendants $P_0(t,T)$. In our proposed importance sampler we consider the probability of extinction $P_1$ of the just created lineage.

Figure 3.5 shows a diagram with the steps of the proposed data augmentation algorithm.

Using equation (3.9) with the data augmentation scheme described above we have the following sampling probability of the full augmentation process:

$$f_m((t,\tau)|x_{obs},\theta) = \prod_{i\,\in\,\mathscr{M}_\tau} f_B(t_i|\ t_{i-1}) P_A(\tau_i|t_i) f_D\left(t_i^e|\tau_i,t_i\right) = \tag{3.11}$$

$$\left[\prod_{i\,\in\,\mathscr{M}_\tau} \frac{\sum_{s\in\mathscr{S}_{t_i}} \lambda_{t_i,s|\theta}}{N^e_{t_i^-}+2N^o_{t_i^-}} \mu_{t_i,s_i^*|\theta} e^{-\mu_{t_i,s_i^*|\theta}\left(t_i^e-t_i\right)}\right] \left[\prod_{i\,\in\,\{1,\ldots,p\}} e^{-\int_{t_{i-1}}^{t_i}\left[\sum_{b\in\mathscr{S}_t}\lambda_{r,b|\theta,v}\left(1-e^{-\mu_{r,b|\theta,v}(t_p-r)}\right)\right]dt}\right] \tag{3.12}$$

Taking the logarithm we have

$$\begin{aligned}\ell_m(\theta) \quad = \quad & -\int_{t_0}^{t_p}\left[\sum_{s\in\mathscr{S}_t}\lambda_{t,s|\theta}\left(1-e^{-\mu_{t,s|\theta}(t_p-t)}\right)\right]dt + \sum_{i\in\mathscr{M}_\tau}\log\left(\sum_{s\in\mathscr{S}_{t_i}}\lambda_{t_i,s|\theta}\left(1-e^{-\mu_{t_i,s|\theta}(t_p-t_i)}\right)\right) \\ & -\log\left(N^e_{t_i^-}+2N^o_{t_i^-}\right)+\log\left(\mu_{t_i,s_i^*|\theta}\right)-\mu_{t_i,s_i^*|\theta}\left(t_i^e-t_i\right)\end{aligned} \tag{3.13}$$

**Example.** Consider a model with a speciation rate that is the same for all lineages and with a constant extinction rate, i.e.,

$$\lambda_{t,s;|\theta} = \lambda_{t|\theta}, \forall\, s\in\mathscr{S}_t, \text{ and } \mu_{t,s;|\theta} = \mu_o, \forall\, t,s;|\theta,$$

then, the sampling probability of the DAA is

$$f_m((t,\tau)|x_{obs},\theta) = \mu_0^{\#\mathscr{M}_\tau}\prod_{i\,\in\,\mathscr{M}_\tau} e^{-\mu_0(t_i^e-t_i)}\frac{N_{t_i^-}\lambda_{t_i|\theta}}{N^e_{t_i^-}+2N^o_{t_i^-}}\prod_{i\,\in\,\{1,\ldots,p\}} e^{-N_{t_i}\int_{t_{i-1}}^{t_i}\left[\lambda_{t|\theta}\left(1-e^{-\mu_0(t_p-t)}\right)\right]dt}.$$

SAMPLE SIZE

In Monte-Carlo methods, the variance of the estimates and the convergence time are determined by the sample size, the explored region of the parameter space and the type of data. From these three factors, we have control only over the sample size. MC methods require a sensible choice of the sample size, and it much depends on the type of problem. In iterative algorithms such as the MCEM algorithm, it is usually efficient to start with small sample size and increase it while parameters are approaching the MLE (Delyon *et al.*, 1999), but there is no general rule for the choice of sampling sizes (Atanassov and Dimov, 2008).

To determine the required sample size in the emphasis method, we consider the estimator the distribution $f_m$.

$$f\,(x_{obs}|\theta) \;=\; \int\limits_{x\in\mathcal{X}(x_{obs})} f(x,x_{obs}|\theta)\,dx \;\approx\; \frac{1}{M}\sum_{x_i\sim f_m}\frac{f(x,x_{obs}|\theta)}{f_m(x_i|x_{obs},\theta)} \;=\; \widehat{f\,(x_{obs}|\theta)}$$

where $\{x_1,...,x_M\}$ are full trees sampled from $f_m((t,\tau)|x_{obs},\theta)$. We will assume that if $SE\left(\widehat{\ell(\theta)}\right) < C$ then $\widehat{\ell(\theta)}$ is good enough, for a small constant $C$. Note that, by taking a Taylor expansion of the logarithm of the estimated likelihood around the observed tree $x_{obs}$, we can write

$$\begin{aligned}\mathbb{E}\left[\log\ \widehat{f\,(x_{obs}|\theta)}\,\right] \;&\approx\; \mathbb{E}[\log f\,(x_{obs}|\theta) + \left(\ \widehat{f\,(x_{obs}|\theta)} - f\,(x_{obs}|\theta)\right)\frac{1}{f\,(x_{obs}|\theta)} \\ &\qquad + \frac{1}{2}\left(\ \widehat{f\,(x_{obs}|\theta)} - f\,(x_{obs}|\theta)\right)^2\frac{-1}{f^2\,(x_{obs}|\theta)}] \\ &=\; \ell(\theta) - \frac{1}{2}\frac{V\left(\widehat{f\,(x_{obs}|\theta)}\right)}{f^2(x_{obs}|\theta)}\end{aligned}$$

where the last term represents the first-order bias. So, typically our method will underestimate the loglikelihood. Furthermore, the estimation tends to be variable. The variability can be assessed by a first order Taylor expansion, i.e.,

$$\begin{aligned}\mathbb{V}\left[\log\ \widehat{f\,(x_{obs}|\theta)}\,\right] \;&\approx\; \mathbb{V}\left[\log f\,(x_{obs}|\theta) + \left(\ \widehat{f\,(x_{obs}|\theta)} - f\,(x_{obs}|\theta)\right)\frac{1}{f\,(x_{obs}|\theta)}\right] \\ &=\; \frac{V\left(\ \widehat{f\,(x_{obs}|\theta)}\right)}{f^2\,(x_{obs}|\theta)}\end{aligned}$$

The variance of $\widehat{f\,(x_{obs}|\theta)}$ can be easily estimated by the sample variance of the importance weights divided by the sample size, i.e., $V\left(\widehat{f\,(x_{obs}|\theta)}\right) \approx V(w_1,...,w_M)/M$. To assess the total possible deviation of the MC estimation we consider the bias and the standard error combined:

$$\widehat{\text{Deviation}} = \frac{1}{2}\frac{V(w_1,...,w_M)}{M\left(\widehat{f\,(x_{obs}|\theta)}\right)^2} + \frac{\sqrt{V(w_1,...,w_M)}}{\sqrt{M}\,\widehat{f\,(x_{obs}|\theta)}}$$

where the weights $w_i$ are given by equation 3.6. The first term of the equation is an estimate of the bias and the second term an estimate of the standard error.

If it is feasible to perform a large number of trial simulations, then the standard error becomes of a lower order than the bias and, thus, we have that approximately,

$$\widehat{\ell(\theta)} \;>\; \ell(\theta) - \widehat{\text{Bias}} \;\approx\; \ell(\theta) - \frac{K_1}{M}$$

for a constant value $K_1 = V(w_1,...,w_M)/\left(2\widehat{f^2\,(x_{obs}|\theta)}\right)$, which can be further used to do a bias-correction of our likelihood. With this, we can calculate an approximation

of the required sample size $M$ to reach a desired level of accuracy. Figure 3.6 shows an illustration on the method we use to assess the required sample size. We sample trees with different sample sizes in order to obtain different estimates of the loglikelihood, and we can fit a curve of the form $c_1 + \frac{c_2}{M}$. With the fitted model we can calculate the asymptotic value of the loglikelihood $c_1$. We set the sample size $M$ such that for a given tolerance level $\epsilon$, $\frac{c_1}{M} < \epsilon$.

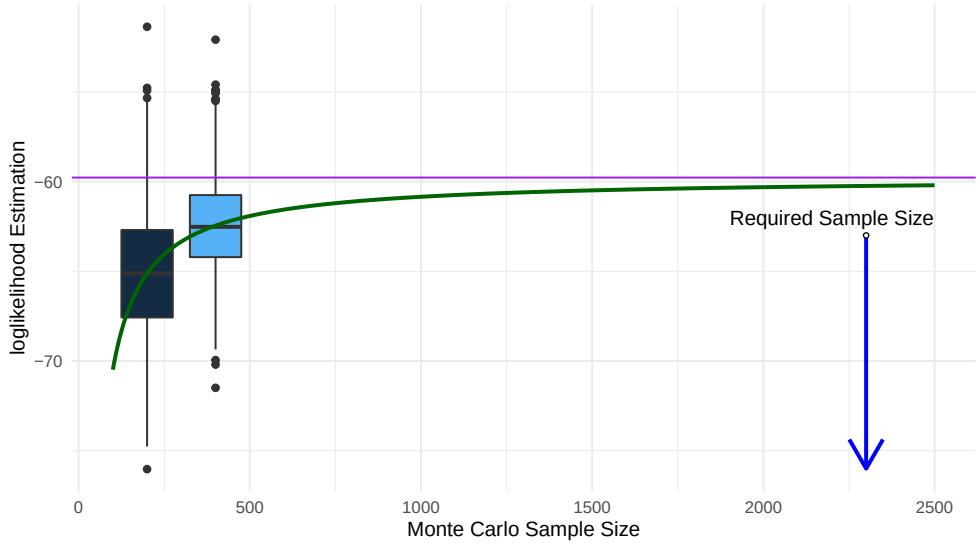



**Figure 3.6 |** To calculate the required sample size, we simulate trees and estimate the loglikelihood via Monte-Carlo with at least two different sample sizes. We then calculate the curve that fits the relationship between the sample size and the estimated MC loglikelihood. This curve indicates the sample size required under a given tolerance level.

The MCEM algorithm can be replaced by the SAEM, MCMC or variations and combinations of them (Delyon *et al.*, 1999; Celeux *et al.*, 1995; Rydén *et al.*, 2008; Wang, 2007; Kuhn and Lavielle, 2004). All these algorithms rely on a sampling scheme and importance samplers. Our sampling scheme and sample size determination strategy can be used in any of these methods.

### 3.3.4. MODEL SELECTION

It is possible to apply standard model selection tools such as AIC or BIC (Wit *et al.*, 2012) to the obtained loglikelihood. Furthermore, in the context of phylogenetic trees, specific statistics have been developed to test how well a model describes an observed tree. An informative summary statistic is the lineage-through-time (LTT) statistic (Janzen *et al.*, 2015), defined as

$$LTT(1,2) = \int_{t_0}^{t_p} |N_t^{(1)} - N_t^{(2)}| dt$$

where $N_t^{(i)}$ is the number of species of a tree $i$ at time $t$. This statistic can also be used to assess how well a model describes an observed tree, simulating trees from the desired model and then calculating the LTT statistic between each simulated tree and the observed tree. It is also possible to calculate the mean number of species through time into a single "average" tree and calculate the LTT statistic of that tree compared to the observed tree.

The LTT statistic and model 3.1 are mathematical expressions that take into account the branching times of the tree, but ignore the topology. That is, the parent-child relationship among species is not relevant; only the branching times are considered. In this chapter, we introduce an alternative to the LTT statistic, considering phylogenetic diversity instead of species richness. We define the *phylodiversity-through-time* (PTT) statistic as

$$PTT(1,2) = \int_{t_0}^{t_p} |P_t^{(1)} - P_t^{(2)}| dt$$

where $P_t^{(i)}$ is the phylogenetic diversity for tree $i$ at time $t$. In Figure 3.7 we present two example trees and the species richness for both trees as well as the phylogenetic diversity. The blue area represents the LTT statistic, while the green area represents the PTT statistic.

In this chapter, we will consider the LTT statistic, the PTT statistic and the AIC weights for model comparison and general goodness-of-fit considerations.

## 3.4. APPLICATION

To illustrate our method, we quantitatively compare model (3.2) with model (3.1) for 14 phylogenies obtained from (Condamine *et al.*, 2019), with sizes ranging between 16 and 141 species and crown ages between 5 My and 65 My. Figure 3.8 represents the distribution of the number of species and crown age of the clades.

In this application, the ultimate use of the emphasis framework is to find the maximum likelihood estimates for the model (3.2) and compare them with model (3.1), to quantify the impact of phylodiversity-dependent diversification. But first, we will perform initial steps to evaluate the required sampling size for different phylogenetic trees. This will give insight about which phylogenies we can apply emphasis to and at what computational cost.

### 3.4.1. MONTE-CARLO APPROXIMATION WITH THE PROPOSED IMPORTANCE SAMPLER

Before performing analysis with the model (3.2), we want to test the efficiency of the Monte-Carlo method with the importance sampler introduced in Section 3.3.3. Monte-Carlo methods require a sensible choice of the sample size, and this largely depends on the type of problem. For sampling full trees, the relationship between accuracy and sample size is complex. For the uniform importance sampler presented in Chapter 2, the required MC sample size becomes huge for most empirical trees. We first want to

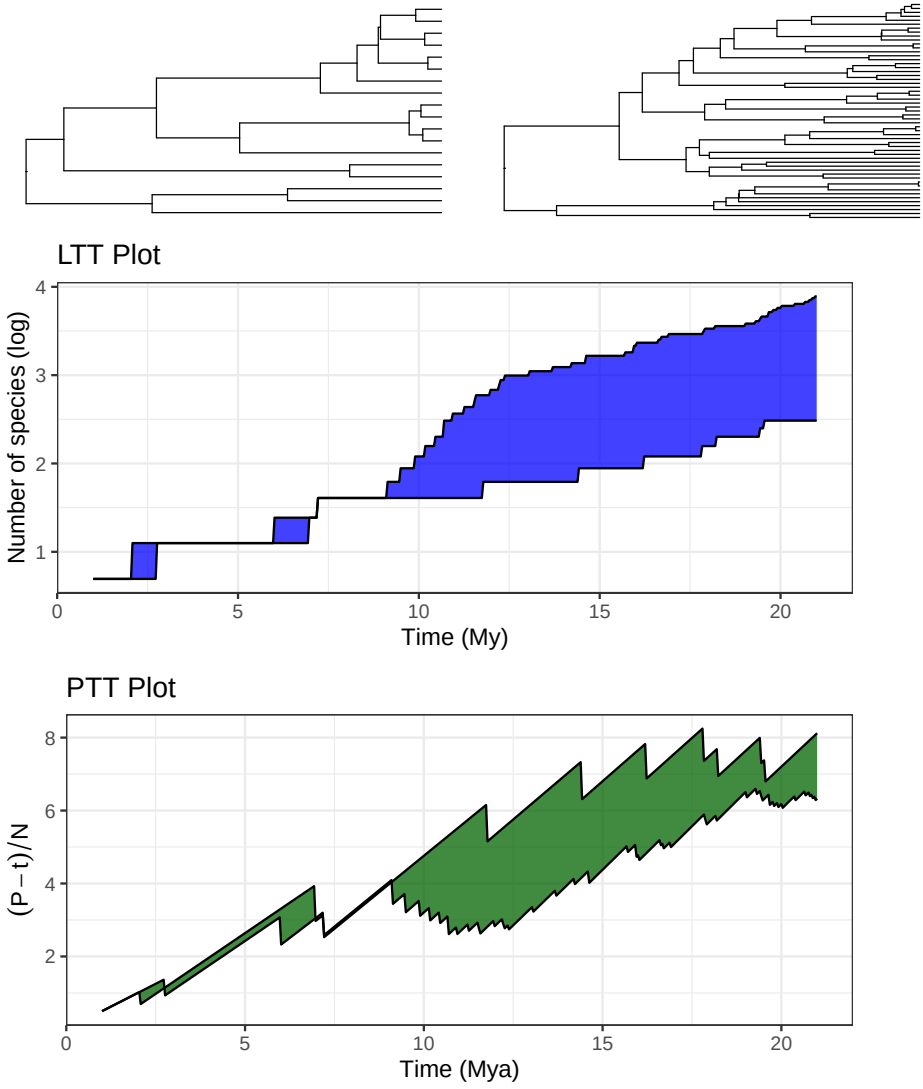
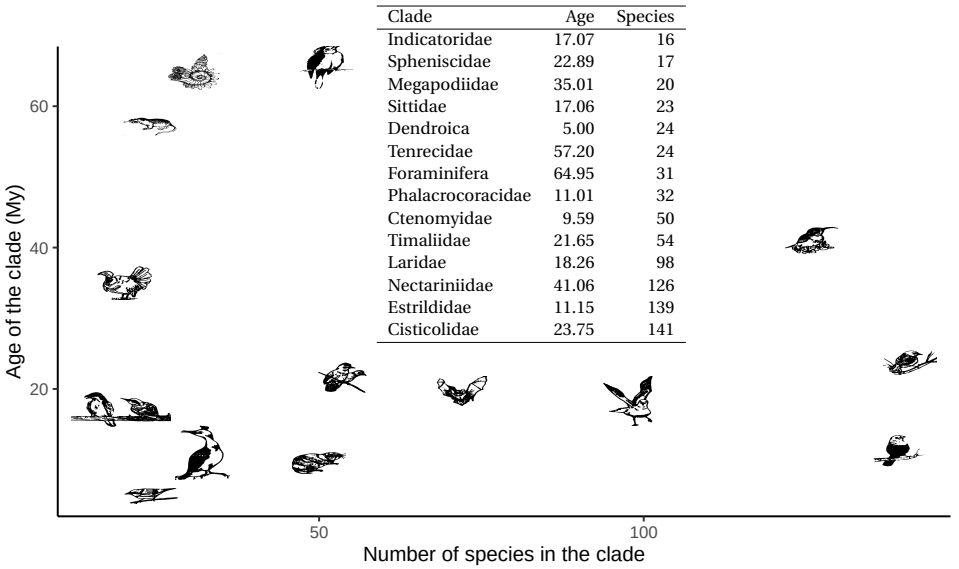



**Figure 3.7 |** Comparison between two phylogenetic trees using the LTT (Number of lineages) and PTT (Phylogenetic diversity) through time. The area represents the distance between the trees.

test that the non-homogenous sampler presented in this chapter can provide accurate approximations.




| Clade | Age | Species |
|---|---|---|
| Indicatoridae | 17.07 | 16 |
| Sphenisciadae | 22.89 | 17 |
| Megapodiidae | 35.01 | 20 |
| Sittidae | 17.06 | 23 |
| Dendroica | 5.00 | 24 |
| Tenrecidae | 57.20 | 24 |
| Foraminifera | 64.95 | 31 |
| Phalacrocoracidae | 11.01 | 32 |
| Ctenomyidae | 9.59 | 50 |
| Timaliidae | 21.65 | 54 |
| Laridae | 18.26 | 98 |
| Nectariniidae | 41.06 | 126 |
| Estrildidae | 11.15 | 139 |
| Cisticolidae | 23.75 | 141 |

**Figure 3.8 |** Distribution of crown ages and numbers of species of 14 phylogenetic trees.

Note that,

$$\begin{aligned}\mathbb{E}\left[f(x_{obs}|\theta)\right] &= \int_{x\in\mathcal{X}(x_{obs})} f(x|\theta)dx \\ &= \int_{x\in\mathcal{X}(x_{obs})} \frac{f(x|\theta)}{f_m(x|\theta,x_{obs})} f_m(x|\theta,x_{obs})dx \\ &\approx \frac{1}{M}\sum_{x_i\sim f_m(x|\theta,x_{obs})} \frac{f(x_i|\theta)}{f_m(x_i|\theta,x_{obs})} \end{aligned} \tag{3.14}$$

so we can use the Monte-Carlo sampling to approximate the likelihood for every parameter. To assess how well our importance sampler does as a function of MC sample size, we compare MC estimations for the 14 phylogenies for the LDD model, for which an existing solution exists. For each phylogeny, we calculate the MLE for the LDD model with the DDD R package. With these parameters, we perform Monte-Carlo sampling and approximate the expectation (3.14) with 4 different MC sampling sizes. In Table 3.1 we show the MC estimations and, in the last column, the analytical solution.

If the difference between the analytical loglikelihood and the MC approximated loglikelihood is less than 1, we will conclude that the estimation is good enough. Under this assumption, we found that a sample size of 1000 is good enough for phylogenies up to approximately 70 species. With a sample size of $10^6$ we have very accurate estimations for all clades with the exceptions of Nectariniidae, Estrildidae and Cisticolidae. These are the larger trees with more than 100 species. Table 3.1 contains detailed estimation for 4 different sample sizes for each phylogeny.

| | $10^2$ | $10^3$ | $10^4$ | $10^5$ | $10^6$ | Analytical |
|---|---|---|---|---|---|---|
| Indicatoridae | -42.04(1.1e-01) | -42.39(1.1e-01) | -41.97(4.9e-02) | -42.01(4.1e-02) | -41.95(4.8e-02) | -41.89 |
| Spheniscidae | -50.22(3.4e-01) | -50.64(1.8e-01) | -50.49(1.1e-01) | -50.23(8.1e-02) | -50.35(4.3e-02) | -50.23 |
| Megapodiidae | -68.98(6.1e-02) | -68.77(4.6e-02) | -68.3(1.9e-02) | -68.1(2.1e-02) | -68.09(1.0e-02) | -68 |
| Sittidae | -64.72(3.8e-01) | -64.19(2.3e-01) | -64.42(1.0e-01) | -64.38(7.6e-02) | -64.32(3.0e-02) | -64.31 |
| dendroica | -38.97(3.5e-01) | -39.2(2.3e-01) | -39.04(1.1e-01) | -38.97(6.0e-02) | -39.03(4.3e-02) | -38.91 |
| Tenrecidae | -89.68(1.7e-01) | -89.12(6.0e-02) | -88.84(3.6e-02) | -89.05(2.1e-02) | -88.42(4.4e-02) | -88.4 |
| foraminifera | -118.48(5.9e-02) | -117.7(4.3e-02) | -116.48(2.4e-02) | -116.34(1.9e-02) | -115.75(1.1e-02) | -115.73 |
| Phalacrocoracidae | -79.92(4.6e-02) | -81.06(9.5e-02) | -80.42(5.7e-02) | -80.46(4.0e-02) | -80.4(2.0e-02) | -80.25 |
| Ctenomyidae | -122.66(1.4e-01) | -120.8(1.1e-01) | -120.61(6.1e-02) | -120.7(4.0e-02) | -120.63(5.6e-02) | -120.65 |
| Timaliidae | -154.23(4.7e-03) | -154.93(5.1e-03) | -153.75(3.0e-03) | -153.91(2.2e-03) | -153.56(9.8e-04) | -153.48 |
| Laridae | -232.12(5.9e-03) | -232.15(3.1e-03) | -231.52(2.0e-03) | -231.23(1.1e-03) | -231.14(8.7e-04) | -231.07 |
| Nectariniidae | -416.2(6.3e-04) | -410.81(2.7e-05) | -404.61(2.0e-06) | -402.19(7.4e-07) | -400.74(3.3e-07) | -399.04 |
| Estrildidae | -321.31(5.4e-03) | -316.31(1.7e-04) | -311.26(2.7e-05) | -312.22(2.8e-05) | -310.93(1.4e-05) | -309.35 |
| Cisticolidae | -410.16(2.5e-04) | -403.73(2.3e-05) | -402.29(1.0e-05) | -402.06(7.3e-06) | -400.6(4.0e-06) | -397.94 |

**Table 3.1 |** Monte-Carlo approximation of the loglikelihood for each tree at its corresponding MLE for the diversification process generated under the LDD model, for different sample sizes. The last column contains the analytical value obtained with the R package DDD.

### 3.4.2. ESTIMATION AND MODEL SELECTION

In Table 3.2 we report the parameter estimations for the two models of interest for the 14 case-study phylogenies. Looking at the LTT statistics, we can observe that there is only a slight improvement of the LPDD model over the LDD model in most of the cases. This is confirmed when we see the loglikelihood values, where the improvement is not large enough to justify preferring the LPDD model, which is confirmed with the AIC weights which always prefer the LDD model.

Note that the AIC weights are based on a Monte-Carlo approximation of the likelihood which will slightly underestimate models with more parameters. As a result, the AIC is more conservative than a test where the AIC values are calculated using the true value of the loglikelihood. Note that in some cases the loglikelihood for the LPDD model is still smaller than the loglikelihood of the LDD model, which cannot be correct because the LDD model is nested within the LPDD model, and hence the LPDD likelihood should be always smaller than the LPDD likelihood. We argue that this is because the Monte-Carlo approximation is not good enough yet. These computational issues suggest that hypothesis testing with AIC might not be an appropriate tool for model selection. Significance tests, instead, do not depend on the approximation of the likelihood but on the approximation of the Hessian of the likelihood (see equation 2.6); because the likelihood is asymptotically quadratic near its maximum (hence the second derivative is constant), the approximation of the Hessian should not present the computational issues that the approximation of the likelihood presents, and hence significance tests seem more reliable. Based on the significance test results, we conclude that the phylodiversity-dependent diversification model provides an alternative/better explanation to/than the diversity-dependent diversification model, at least in some of our clades.

In Figure 3.9 we see an example of the expected lineages-through-time plot for each model in comparison with the observed lineage through time plot corresponding to the Timaliidae phylogeny, and the speciation rates through time plot. We can see that both models agree that speciation happened roughly at a rate of 0.2 species per million rate during the last 10 million years; however, they diverge on the estimates for the

period between 20 and 10 million years ago. Including phylogenetic diversity involves a fluctuating speciation rate around 0.2 spe/Mye reaching its maximum around 15 years ago while the LDD model assumes a monotonously decreasing speciation rate. In general, the difference between the two models is not large and this pattern is present across all the 14 phylogenies.

Finally, in Table 3.3 we report the loglikelihood estimates for the LPDD model. We see that for most of the cases the sample size was large enough, but for larger trees the convergence performs much slower for the LPDD case than for the LDD case.

| Clade | Age | Tips | Model | AICw | LTT | PTT | loglikelihood | $\mu_0$ | $\lambda_0$ | $\beta_N$ | $\beta_P$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Indicatoridae** | 17.07 | 16 | LDD | 0.73 | 0.49 | 0.49 | -42.01 | 0.22 | 1.62 | -0.085 | 0 |
| | | | LPDD | 0.27 | 0.51 | 0.51 | -41.99 (6e-02) | 0.21 (4e-03) | 1.61 (5e-04) | -0.085 (5e-05) | -0.001 (3e-04) |
| **Spheniscidae** | 22.89 | 17 | LDD | 0.83 | 0.41 | 0.52 | -50.23 | 0.2 | 1.61 | -0.081 | 0 |
| | | | LPDD | 0.17 | 0.59 | 0.48 | -50.79 (3e-02) | 0.16 (2e-03) | 1.49 (1e-02) | -0.079 (6e-04) | 0.004 (7e-04) |
| **Megapodiidae** | 35.01 | 20 | LDD | 0.78 | 0.45 | 0.48 | -68.1 | 0.1 | 0.83 | -0.036 | 0 |
| | | | LPDD | 0.22 | 0.55 | 0.52 | -68.37 (3e-02) | 0.09 (1e-03) | 0.83 (6e-04) | -0.036 (4e-05) | 0 (1e-04) |
| **Sittidae** | 17.06 | 23 | LDD | 0.79 | 0.5 | 0.46 | -64.38 | 0.15 | 0.58 | -0.018 | 0 |
| | | | LPDD | 0.21 | 0.5 | 0.54 | -64.72 (1e-01) | 0.12 (9e-04) | 0.4 (1e-03) | -0.021 (3e-04) | 0.039 (1e-03) |
| **Dendroica** | 5.00 | 24 | LDD | 0.77 | 0.46 | 0.53 | -38.97 | 0.16 | 3.05 | -0.117 | 0 |
| | | | LPDD | 0.23 | 0.54 | 0.47 | -39.16 (8e-02) | 0.14 (1e-03) | 2.99 (2e-02) | -0.118 (1e-03) | 0.007 (3e-03) |
| **Tenrecidae** | 57.20 | 24 | LDD | 0.74 | 0.46 | 0.47 | -89.05 | 0.11 | 0.59 | -0.02 | 0 |
| | | | LPDD | 0.26 | 0.54 | 0.53 | -89.1 (3e-02) | 0.09 (6e-04) | 0.59 (4e-04) | -0.02 (8e-05) | 0.001 (2e-04) |
| **Foraminifera** | 64.95 | 31 | LDD | 0.84 | 0.39 | 0.42 | -116.34 | 0.1 | 1.18 | -0.034 | 0 |
| | | | LPDD | 0.16 | 0.61 | 0.58 | -117.01 (4e-03) | 0.08 (4e-04) | 1.17 (1e-04) | -0.034 (2e-06) | 0 (4e-06) |
| **Phalacrocoracidae** | 11.01 | 32 | LDD | 0.71 | 0.41 | 0.48 | -80.46 | 0.24 | 1.67 | -0.044 | 0 |
| | | | LPDD | 0.29 | 0.59 | 0.52 | -80.34 (4e-02) | 0.24 (3e-03) | 1.59 (2e-02) | -0.038 (3e-04) | -0.027 (3e-03) |
| **Ctenomyidae** | 9.59 | 50 | LDD | 0.74 | 0.46 | 0.42 | -120.7 | 0.16 | 1.15 | -0.02 | 0 |
| | | | LPDD | 0.26 | 0.54 | 0.58 | -120.75 (6e-02) | 0.14 (1e-03) | 1.08 (6e-03) | -0.019 (2e-04) | 0.007 (2e-03) |
| **Timaliidae** | 21.65 | 54 | LDD | 1 | 0.48 | 0.53 | -153.91 | 0.14 | 0.5 | -0.006 | 0 |
| | | | LPDD | 0 | 0.52 | 0.47 | -158.63 (7e-03) | 0.07 (1e-03) | 0.22 (8e-03) | -0.006 (2e-04) | 0.034 (2e-03) |
| **Laridae** | 18.26 | 98 | LDD | 0.68 | 0.19 | 0.5 | -231.23 | 0.13 | 0.32 | 0 | 0 |
| | | | LPDD | 0.32 | 0.81 | 0.5 | -231.01 (6e-02) | 0.02 (1e-03) | 0.46 (2e-03) | 0.001 (2e-05) | -0.061 (7e-04) |
| **Nectariniidae** | 41.06 | 126 | LDD | 0.66 | 0.54 | 0.83 | -402.19 | 0.14 | 0.32 | -0.001 | 0 |
| | | | LPDD | 0.34 | 0.46 | 0.17 | -401.83 (4e-02) | 0.02 (4e-04) | 0.25 (1e-03) | 0 (2e-05) | -0.019 (3e-04) |
| **Estrildidae** | 11.15 | 139 | LDD | 0.88 | 0.5 | 0.77 | -312.22 | 0.28 | 1.05 | -0.005 | 0 |
| | | | LPDD | 0.12 | 0.5 | 0.23 | -313.21 (6e-03) | 0.12 (2e-03) | 0.42 (7e-03) | -0.006 (3e-04) | 0.176 (1e-02) |
| **Cisticolidae** | 23.75 | 141 | LDD | 0.51 | 0.48 | 0.76 | -402.06 | 0.16 | 0.48 | -0.002 | 0 |
| | | | LPDD | 0.49 | 0.52 | 0.24 | -401.11 (2e-02) | 0.05 (1e-03) | 0.37 (4e-03) | 0 (5e-05) | -0.04 (1e-03) |

**Table 3.2 |** Parameter estimations for LDD and LPDD model for 14 phylogenies. The fifth column shows the AIC weights for the comparison of these two models. The sixth column is the normalised LTT statistic. The last four columns represent the parameter estimates. Between parentheses we report the standard deviation of the Monte-Carlo approximation.

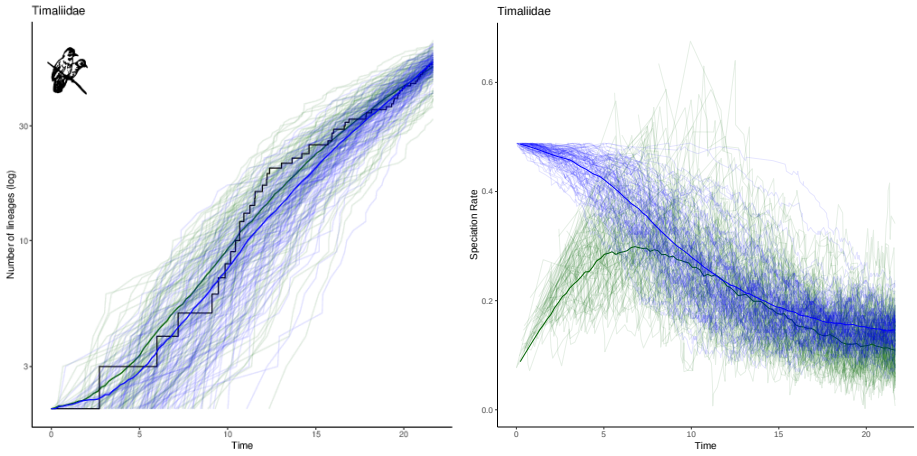
## 3.5. DISCUSSION

Diversity-dependent diversification models have been developed during the last decade in order to understand and quantify the existence and impact of ecological limits to macroevolutionary dynamics. At the moment, only models with a dependence of diversification rates on species richness have been implemented, but these models ignore other facets of diversity, such as phylodiversity.

In this chapter, we have completed the statistical methodology introduced in Chapter 2, with the design of a data augmentation scheme that provides an efficient importance

| | $10^2$ | $10^3$ | $10^4$ | $10^5$ | $10^6$ |
|---|---|---|---|---|---|
| Indicatoridae | -42.49(1.3e-01) | -42.15(1.0e-01) | -42.26(6.5e-02) | -41.99(6.1e-02) | -42.02(2.7e-02) |
| Spheniscidae | -51.66(2.4e-01) | -50.68(1.1e-01) | -50.9(6.4e-02) | -50.79(3.0e-02) | -50.75(1.6e-02) |
| Megapodiidae | -68.97(2.1e-01) | -68.19(8.1e-02) | -68.42(6.2e-02) | -68.37(3.5e-02) | -68.26(4.0e-02) |
| Sittidae | -65.49(2.7e-01) | -64.49(2.0e-01) | -64.83(1.2e-01) | -64.72(1.2e-01) | -64.73(2.4e-02) |
| dendroica | -39.25(2.9e-01) | -39.33(1.9e-01) | -39.22(9.4e-02) | -39.16(8.0e-02) | -39.14(3.2e-02) |
| Tenrecidae | -89.06(8.6e-02) | -88.65(6.0e-02) | -89.61(3.9e-02) | -89.1(3.4e-02) | -89.12(1.5e-02) |
| foraminifera | -119.3(1.2e-02) | -117.71(5.1e-03) | -117.94(3.4e-03) | -117.01(4.5e-03) | -117.3(2.3e-03) |
| Phalacrocoracidae | -79.61(1.2e-01) | -81.02(8.5e-02) | -80.54(6.5e-02) | -80.34(4.1e-02) | -80.56(2.1e-02) |
| Ctenomyidae | -119.95(1.8e-01) | -121.08(2.0e-01) | -120.87(8.2e-02) | -120.75(5.6e-02) | -120.77(7.6e-02) |
| Timaliidae | -159.57(4.8e-02) | -158.95(2.0e-02) | -158.37(1.4e-02) | -158.63(6.8e-03) | -158.16(6.9e-03) |
| Laridae | -230.98(6.3e-01) | -230.86(2.2e-01) | -231.03(1.0e-01) | -231.01(5.6e-02) | -231(3.8e-02) |
| Nectariniidae | -402.09(1.9e-01) | -402.01(9.6e-02) | -401.81(6.9e-02) | -401.83(3.9e-02) | -401.87(3.0e-02) |
| Estrildidae | -315.33(3.0e-02) | -313.25(1.3e-02) | -313.13(1.1e-02) | -313.21(6.0e-03) | -313.28(3.7e-03) |
| Cisticolidae | -403.78(5.9e-02) | -401.9(3.6e-02) | -401.52(2.4e-02) | -401.11(1.7e-02) | -401(8.2e-03) |

**Table 3.3 |** Loglikelihood approximations of the LPD model at its MLE value for the 14 phylogenies.




**(a)** Lineages through time

**(b)** Speciation rate through time

**Figure 3.9 |** Evolution of extant species richness (LTT-plot) and evolution of global speciation rates for 13 clades under LDD (blue) and LPD (green) models.

sampler, which is a substantial improvement in comparison to the uniform importance sampler considered in the Chapter 2, as it enables applying the method to a large number of empirical phylogenies.

In the application to 14 example phylogenies, we studied the LPDD model, i.e., a model with a linear effect of phylodiversity on speciation. We found that including phylodiversity does not provide a substantial improvement in comparison with richness-dependent diversification models. However, phylodiversity does provide an alternative and slightly more complete explanation to speciation dynamics; the LTT statistic and the PTT statistic provide insights and, most of the times, reflects that trees generated by the LPDD model are closer to real phylogenies than the trees generated under the LDD model. While the model with fewer parameters is preferred using AIC, the phylodiversity component is statistically significant, suggesting that it should not be ignored.

This may not be the final word because there are some technical improvements to be made. In particular, we did not condition the likelihood on non-extinction of the clade; even though this is generally recommended (Etienne *et al.*, 2016; Stadler, 2013). Such conditioning is covered in Chapter 5.

Our method is not limited to phylogenetic diversity-dependent diversification models, but allows inference of a general class of species diversification models, considering time, traits, climate, functional diversity, just to name a few. With the data augmentation described her we have provided a general tool that can be potentially used to quantify and test a large number of hypotheses in macroevolutionary diversification.

# 4

# LINEAGE-DEPENDENT PHYLOGENETIC DIVERSITY AS A DRIVER OF SPECIES DIVERSIFICATION

*The 'art' of building a good model is to capture the essential features of the biology without burdening the model with non-essential details.*

Darren J. Wilkinson

Richter F, Wit EC, Etienne RS. In preparation.

## ABSTRACT

*Modelling species diversification processes and performing statistical inference on these processes using phylogenetic trees is an active area of research. It requires the development of novel quantitative tools to study the influence of ecological factors on (macro)evolutionary processes. The Yule model or constant-rate birth-death models are still widely used, because of their simplicity that allows fast computation of the likelihood, i.e., the probability of the phylogenetic tree given the diversification model parameters.*

*The development of more complex species diversification models that consider additional factors typically involves the computation of the likelihood for the diversification model, via the master equation. These more complex models consider species interactions and need to integrate across all such interactions, due to the lack of information about these interactions in the past — which are unlikely to ever become available.*

*A promising alternative is to use a proxy for (past) species interactions. Species diversity is one such proxy, and it has indeed been possible to compute the likelihood for models in which diversification rates depend on diversity. However, this proxy assumes that all species interact in the same way. To accommodate variation in these interactions, we propose to use phylogenetic diversity as a proxy, because phylogenetic diversity between species, defined as the time to the most common ancestor, represents the niche distance among species.*

*In this chapter, we integrate per-species phylogenetic distance into diversity-dependent diversification models, the results of which we will call lineage-dependent diversification models. We show that these models cover a broad range of topologies, consistent with those of real trees. In addition, we develop a stochastic gradient descent framework that will enable parameter estimation for these models. In summary, with the minimal modification of phylodiversity dependence we expand diversity-dependent diversification models to represent a much broader range of models that can mimic complex topological characteristics of phylogenetic trees.*

## 4.1. INTRODUCTION

Studying the mechanisms underlying species diversification has been an active area of research (Ragan, 2009) and particularly during the last three decades since large-scale DNA sequencing and phylogenetic analysis has been possible (Reynolds, 1973; Nee *et al.*, 1994; Morlon, 2014). Larger and more accurate phylogenies continue to appear (Jetz *et al.*, 2012; Upham *et al.*, 2019; Ramírez-Barahona *et al.*, 2020; Condamine *et al.*, 2019; Hedges *et al.*, 2015b) and species diversification models are more and more sophisticated in order to capture and study multiple hypotheses on how species diversified (Morlon, 2014; Ricklefs, 2007; Etienne and Apol, 2009). In 1925 Yule published a mathematical characterisation of a process where species diversifies with a constant rate without extinction (Yule, 1925). In 1948, Kendall generalised Yule's results by allowing for extinction and time dependent speciation and extinction rates (Kendall *et al.*, 1948). In 1994 Nee et al. presented the likelihood for the time-dependent birth-death process given a phylogenetic tree (Nee *et al.*, 1994), which typically does not contain extinct species. In the last 20 years, a large number of species diversification models have been developed, including diversity-dependent (Etienne *et al.*, 2012b), state-dependent (Maddison *et al.*, 2007; Herrera-Alsina *et al.*, 2019; FitzJohn *et al.*, 2009; Paradis, 2008; Ng and Smith, 2014), and (paleo-)environment-dependent (Condamine *et al.*, 2019; Lewitus and Morlon, 2017) diversification rates. Still, these models have only scratched the surface of all possible diversification processes and more inference methods are needed (Rabosky and Goldberg, 2015).

Maximum likelihood approaches have become a standard to compare various macroevolutionary scenarios using reconstructed phylogenies (Nee, 2006), even though this comparison may have limitations on identifiability (Louca and Pennell, 2020). The design of stochastic birth-death-type species diversification models (SDM) lends itself well for easy testing of hypotheses. Within SDM we can identify two nested classes of models. One class considers global diversification rates, i.e. all species have the same probability to speciate or become extinct. A more general class of SDM considers diversification rates that can differ between species. We will call these models lineage-dependent diversification (LDD) models. Models that assume a global rate for all lineages (lineage-independent diversification (LID) models) are by far the most used and are generally assumed to be a good starting point for analysis. Current LDD diversification models range from simply assuming a shift in the rates (Laudanno *et al.*, 2020c; Rabosky, 2014; Höhna *et al.*, 2019; Maliet *et al.*, 2019), or dependence on a dynamic state (Maddison *et al.*, 2007).

Despite the development of sophisticated (LDD) models, simple constant-rate birth-death models are still commonly used, even though their predictions on temporal (Phillimore and Price, 2008) and topological (Heard, 1996; Mooers *et al.*, 2007; Purvis *et al.*, 2011; Shao, 1990) properties deviate from those in empirical phylogenies. One of the reasons for the limited use of LDD models is that likelihood calculation is much more complicated (Laudanno *et al.*, 2020b) than for lineage-independent diversification (LID) models. However, LDD models are the next generation models that are needed to incorporate more complex ecological interactions, such as niche differentiation and/or facilitative interactions (Barraclough, 2015; Fox, 2005; Olave *et al.*, 2020; Bairey *et al.*, 2016; Roy *et al.*, 2020). Current diversity-dependent diversification models consider such interactions by

simply accounting for the role of species richness on diversification, and for more than a decade diversity-dependence models have already been extensively used in macroevolutionary analysis, detecting clade-level "carrying capacities" and studying the influence of species richness on macroevolutionary processes. In the previous chapter, we generalised diversity-dependent diversification models by allowing diversification rates to depend on phylogenetic diversity, and hence not only the number of species but also their distinctiveness is taken into account. However, this model still assumes that all species are equally likely to diversify, and is therefore an LID model. Current inference procedures mostly use the branching times of the trees as their only input. With LDD models we can take into account topology as well. Per-species phylogenetic distance, defined as the time of the most common ancestor among two species, has not been included in phylogenetic analysis for macroevolutionary studies while it serves as one of the most common proxies for ecological similarities.

In this manuscript, we present and study a LDD model, the lineage-dependent phylodiversity-dependent (LDPD) model and study its effect on macroevolutionary processes. First, in Section 4.2, we discuss the relationship between tree shape and evolutionary advantage, the state of the art of LDD models, and current challenges and we propose a general LDD model that satisfies several desired biological and mathematical properties that makes it a powerful tool for quantitative macroecological and phylogenetic analysis. Then, in Section 4.3, we introduce the lineage-dependent phylodiversity-dependent models and analyse how these models can help capture proper tree shapes and balance levels observed in current phylogenies. We describe the Phylodiversity Matrix, as a dynamical matrix that captures the genetic distance among pairs of species. Finally, in section 4.4 we provide a methodology (stochastic gradient descent) for parameter estimation and derive the required equations for the LDPD model. We use the data augmentation algorithm introduced in the previous chapter to approximate the gradient of the likelihood. We discuss potential directions, advantages and limitations of the method.

## 4.2. MODE AND TEMPO IN EVOLUTIONARY PROCESSES AND REAL PHYLOGENIES

Mathematically the diversification process is characterised by two components, time and balance. Biologically, they represent the tempo and mode of the macro-evolutionary dynamics. Several statistics or measurements have been designed to describe both components. The gamma index describes the distribution of the waiting times between events throughout the process (Pybus and Harvey, 2000; Fordyce, 2010). It is especially useful to capture diversification rate decreases compared with higher rates in the past. The $\rho$-metric introduced in Pigot *et al.* (2010) is an alternative to the gamma-statistic providing values between -1 and 1 indicating speedup in speciation rates towards 1. Regarding the topology of the tree, the Colless index (Colless, 1982) is probably the most used statistic for characterising tree balance; however, dozens of other indices to summarise the shape of the tree have been developed. One example is the Sackin's index (M. Coronado *et al.*, 2020), which computes the sum of the number of ancestors for each tip of the tree; Another example is the Cophenetic index that computes the sum of the

depths of the least common ancestor of every pair of leaves of the tree. The Cophenetic index is a proportion of the MPD index in (Mazel *et al.*, 2016), which calculates the mean pairwise distance between all species in the clade, that is, the mean of the per-species phylogenetic diversity. In this manuscript, we focus on Gamma and Colless, but our analysis can be easily replicated with other metrics.

In figure 4.1, we show the distribution of tempo and mode of 64 Mammal phylogenies represented by the Gamma (GI) and Colless index (CI). We use a normalised version of the CI (PDA normalisation), where high values represent highly unbalanced trees and values close to zero represent highly balanced trees. We use the R packages `ape` and `treeshape` to calculate gamma and Colless indices respectively.

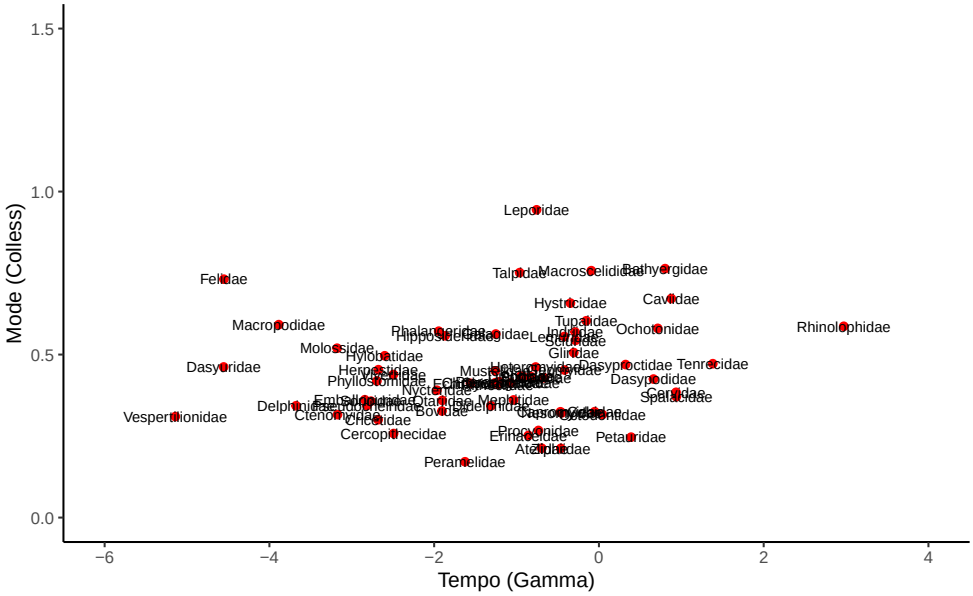



**Figure 4.1 |** Distribution of tempo and mode for 64 mammal phylogenies, characterised by the Gamma and Colless index.

Many studies have analysed and reported gamma and Colless indexes for standard species diversification models (Mooers and Heard, 1997), but this has not been done extensively for diversity-dependent diversification models. We performed a simulation study to study how well such models capture these tree features. We used the standard linear diversity dependence model (LDD) (Etienne *et al.*, 2012b) for four sets of parameters. In Figure 4.2, we plot the four extreme phylogenies with maximum and minimum indices. To be consistent with observed phylogenies, we computed the MLE for the four extreme phylogenies described above, and used these to simulate phylogenies under the LDD model and computed the Gamma and Colless distributions. Figure 4.3 shows that both extremes in the mode (Colless) and one in the tempo (gamma) cannot be mimicked by the simulations.

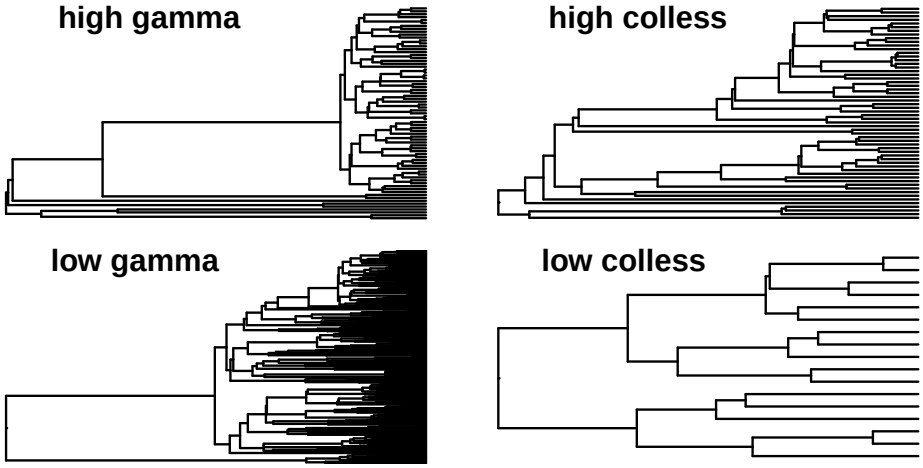



**Figure 4.2 |** Example of the 4 most extreme mammal phylogenies. Rhinolophidae is the clade with maximum gamma index, Leporidae is the clade with maximum Colless index, Vespertilionidae is the clade with minimum gamma index and Peramelidae is the clade with minimum Colless index.

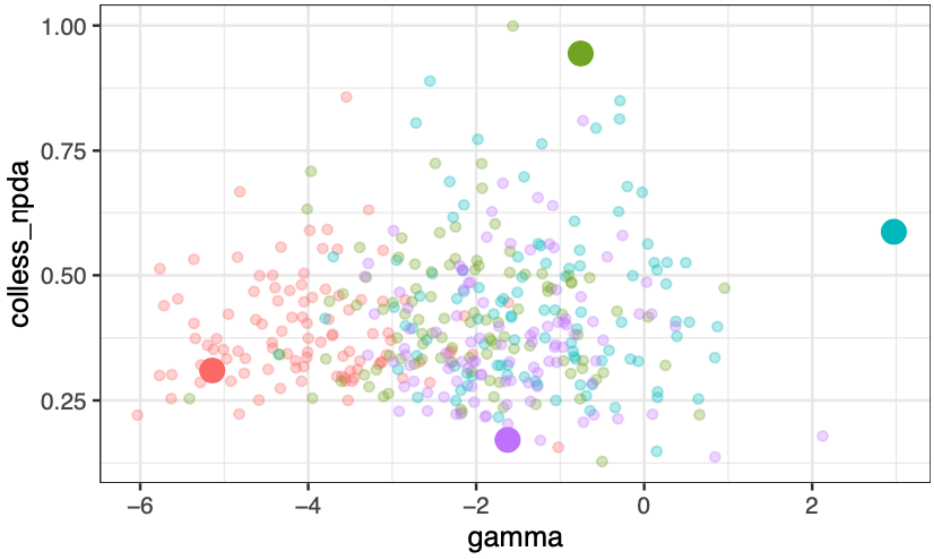



**Figure 4.3 |** Comparison of LDD simulations with empirical phylogenies. Each small point represents the Colless vs gamma coordinate of a simulated tree under the LDD model. Each colour represents a parameter combination for which we used the maximum likelihood estimators for the four trees shown in Figure 4.2. The Colless and Gamma statistics for these four trees are shown as large circles.

## 4.3. THE PHYLOGENETIC-DIVERSITY MATRIX IN LID MODELS

The development of LDD models has only just started; the vast majority of developed and used SDM are LID models (Morlon, 2014), especially because if their computational simplicity (Slowinski and Guyer, 1989). The few LDD models (e.g. (Oliveira *et al.*, 2020)) do not take into account ecological interactions and other essential properties of the diversification process. Here we aim to generalise diversity-dependent diversification models in order to keep them flexible enough to capture mode and tempo, even in extreme phylogenetic trees. We thus search for a model which: (1) incorporate time-varying carrying capacities (Marshall and Quental, 2016), (2) has heritable rates (Caron and Pie, 2020), (3) has the flexibility to promote speciation for younger species for unbalanced trees and promotes speciation in older clades for balanced trees (Jones, 2011), (4) considers community interactions among lineages, (5) considers the dynamical nature of niche diversity (Smaldino *et al.*, 2019) as the ecological role that an organism plays in an ecosystem changes.

### 4.3.1. PHYLOGENETIC DIVERSITY

Phylogenetic diversity or phylodiversity is an ideal candidate for capturing interactions between species, because it is associated with functional diversity (Oliveira *et al.*, 2020), character diversity and other ecological features, although there is still some controversy about this association (Mazel *et al.*, 2018; Tucker *et al.*, 2016). Here we use a per-species phylogenetic diversity index, so we can model LID models where the speciation rate of each species is proportional to the phylogenetic distance of this species holds to the rest of the species in the clade. For this purpose we define a phylogenetic diversity matrix, known also as phylogenetic distance matrix although this term is not only restricted in the literature to the process here defined.

Let $\mathscr{S}_t$ be the set of all species in the phylogenetic tree at time $t$. We define the phylogenetic diversity matrix $P(t)$ as a dynamic matrix, with dynamic dimension $|\mathscr{S}_t| \times |\mathscr{S}_t|$, that takes into account the phylogenetic distance between species. The entries of the matrix are defined as the times to the most recent common ancestor for each pair of species,

$$P_{ij}(t) = \text{time to most recent common ancestor of species } i \text{ and } j.$$

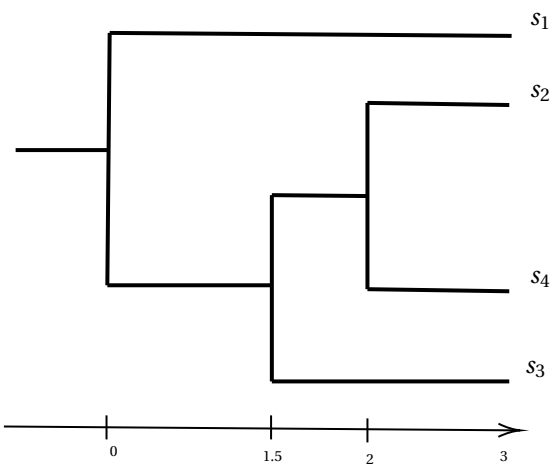
Figure 4.4 shows a simple tree as an example with calculations of the phylogenetic diversity matrix at three different times.

We then define for each species $s$ the mean phylogenetic diversity (Mazel *et al.*, 2016) as

$$P_{s,t} = \frac{1}{|\mathscr{S}_t|} \sum_{s' \in \mathscr{S}_t} P_{s,s'}(t);$$

which is proven to be closely related to Faith's phylogenetic diversity (Faith, 1992) which is widely used in both macroevolution and ecology. We define the overall mean phylogenetic diversity as

$$PDM_t = \frac{1}{|\mathscr{S}_t|} \sum_{s \in \mathscr{S}_t} P_{s,t}$$




$P(1) =$

| | $s_1$ | $s_2$ |
|---|---|---|
| $s_1$ | 0 | 1.1 |
| $s_2$ | 1.1 | 0 |

$P(1.6) =$

| | $s_1$ | $s_2$ | $s_3$ |
|---|---|---|---|
| $s_1$ | 0 | 1.6 | 1.6 |
| $s_2$ | 1.6 | 0 | 0.1 |
| $s_3$ | 1.6 | 0.1 | 0 |

$P(2.5) =$

| | $s_1$ | $s_2$ | $s_3$ | $s_4$ |
|---|---|---|---|---|
| $s_1$ | 0 | 3 | 3 | 3 |
| $s_2$ | 3 | 0 | 1.5 | 1 |
| $s_3$ | 3 | 1.5 | 0 | 1.5 |
| $s_4$ | 3 | 1 | 1.5 | 0 |

**Figure 4.4 |** Phylogenetic diversity matrix at three different time points

which represents the average distance that each species has with the rest of the species in the phylogeny. It describes how phenotypically distinct it is from the other extant species.

These quantities have both a robust biological meaning and elegant and convenient mathematical properties. Note that, between branching times (i.e in periods when no speciation or extinction happens) we have that

$$P_{s,t_i+t} = P_{s,t_i} + t; \qquad PDM_{t_i+t} = PDM_t + t \tag{4.1}$$

Our definition of $PDM_t$ entails

$$\sum_s \left(PDM_t - P_{t,s}\right) = 0 \tag{4.2}$$

We use these properties to develop fast and efficient inference algorithms.

### 4.3.2. THE LID MODELS

We here propose a generalisation of the diversity dependence model (Etienne *et al.*, 2012b) which considers differences among species introduced in the speciation rate,

$$\lambda_{t,s} = \lambda_0 + \beta_N(2 - N_t) + \beta_P\left(PDM_t - P_{t,s}\right); \qquad \lambda_0 > 0, \beta_N > 0, \beta_P > 0 \tag{4.3}$$

This model considers a speciation rate that linearly decreases with the number of species as in the usual LDD model (Etienne *et al.*, 2012b) and adds a LID effect which gives a speciation advantage to species with on average shorter distance to other species. This model, called the lineage-dependent phylodiversity-dependent diversification (LDPD) model thus assumes that species that speciate faster will produce species that speciate fast

as well (Caron and Pie, 2020), which will lead to more unbalanced trees. If we change the constraint $\beta_P > 0$ to $\beta_P < 0$, the model assumes that species that are more phylogenetically distant to other species are more likely to speciate (Nyman, 2010), resulting in more balanced trees.

Note that the properties (4.1, 4.2) imply that the overall speciation rate does not accelerate or decrease relative to the LDD overall speciation rate but only creates differences between the rates of different species.

To analyse how the LDPD model can capture balance and tempo we performed a simulation study. We fixed parameters $\lambda_0 = 0.5, \beta_N = -0.05, \mu_0 = 0.1$ and we varied $\beta_P = 0, 0.001, 0.05, 0.1, 0.5$. We performed 100 simulations for each parameter value. Figure 4.5 shows the distribution of mode and tempo of the simulated data. We can see that by changing $\beta_P$ we can cover the different balance values found in empirical trees, while the distribution of gamma remains wide. This simulation shows the flexibility of the model presented here, especially in relationship with topology. Note that the scale in both indices is larger than for the mammal phylogenies, which suggest that with this model, we can cover balance and tempo observed in nature.

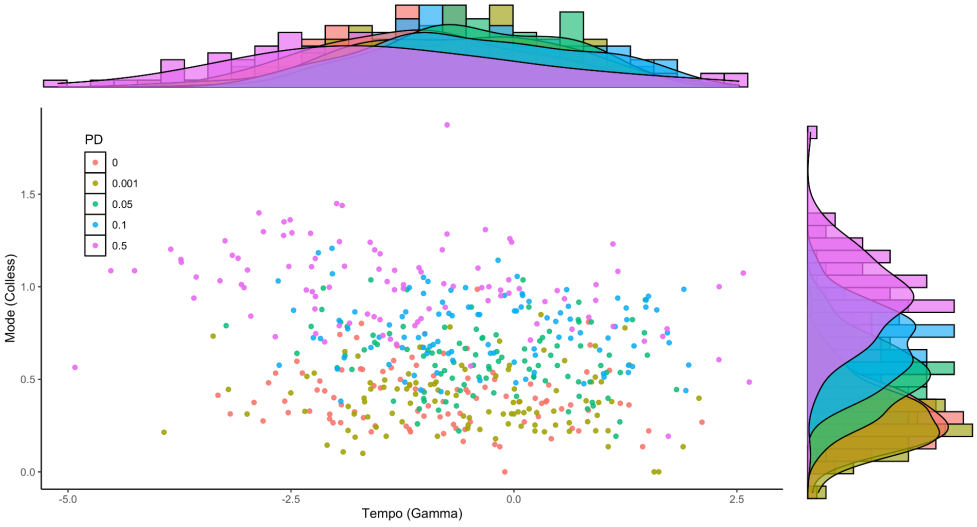



**Figure 4.5 |** Distribution of tempo and mode, characterised by the Gamma and Colless index, for 5 different values for the PD effect. We used $\lambda_0 = 0.5, \beta_N = -0.05, \mu_0 = 0.1$ and we constrain simulation to a crown time of 15My.

## 4.4. PARAMETER ESTIMATION

In previous chapters we developed an MCEM algorithm for likelihood optimisation. In this chapter, we propose a different approach where the EM optimisation is replaced by a stochastic gradient descent method (Robbins and Monro, 1951; Chen *et al.*, 2014).

The aim is to maximise the likelihood function

$$f(y|\theta) = \int_{x\in\mathscr{X}(y)} f(x,y|\theta)\,dx = \int_{x\in\mathscr{X}(y)} \frac{f(x,y|\theta)}{g_\theta(x)} g_\theta(x)dx = \mathbb{E}_{x\sim g_\theta}\left[\frac{f(x,y|\theta)}{g_\theta(x)}\right]$$

where $y$ is the observed phylogenetic tree and $x$ is a variable describing all full trees that are in agreement with $y$. The distribution or importance sampler $g_\theta$ can be, for instance, the uniform sampler introduced in Chapter 2 or the efficient emphasis algorithm developed in Section 3.3.3.

Thus, the maximum likelihood estimator is

$$\hat{\theta} = \arg\max_\theta \mathbb{E}_{x\sim g_\theta}\left[\frac{f(x,y|\theta)}{g_\theta(x)}\right]$$

To maximise this function, we propose a Stochastic Gradient Descent (SGD), which iteratively computes

$$\theta_i = \theta_{i-1} - \eta\, G(\theta)$$

where $\eta$ is a step size (also known as the learning rate in the machine learning literature) and $G(\theta)$ is the gradient of the likelihood function

$$\begin{aligned} G(\theta) &= \nabla\, \mathbb{E}_{x\sim g_\theta}\left[\frac{f(x,y|\theta)}{g_\theta(x)}\right] \\ &= E_{x\sim g_\theta}\frac{f(x,y|\theta)}{g_\theta(x)}\left[\frac{\partial \log f(x,y|\theta)}{\partial\theta} - \frac{\partial \log g_\theta(x)}{\partial\theta}\right] \end{aligned}$$

This gradient can typically not be calculated analytically, but we can use an unbiased Monte-Carlo estimator,

$$\widehat{G(\theta)} \approx \frac{1}{n}\sum_{x_i\sim g_\theta} \frac{f(x_i,y|\theta)}{g_\theta(x_i)}\left[\frac{\partial \log f(x_i,y|\theta)}{\partial\theta} - \frac{\partial \log g_\theta(x_i)}{\partial\theta}\right],$$

where $n$ is the number of sampled trees from the DAA developed in section 3.3.3. Thus, we compute the next-step iteration of the SGD as

$$\theta_i = \theta_{i-1} - \eta\,\widehat{G(\theta)}.$$

This can be evaluated by observing that in the case of species diversification processes, the loglikelihood function is

$$\log f(x|\theta) = \sum_{i\in\mathscr{H}_{spe}} \log\left(\lambda_{t_i,s_i^*|\theta}\right) + \sum_{i\in\mathscr{H}_{ext}} \log\left(\mu_{t_i,s_i^*|\theta}\right) - \int_{t_0}^{t_p}\sum_{s\in\mathscr{S}_t}(\lambda_{t,s|\theta}+\mu_{t,s|\theta})dt$$

where $\mathscr{H}_{spe}$ is the set of indices where the $i$-th event is speciation and $\mathscr{H}_{ext}$ is the set of indices where the $i$-th event is an extinction, $t_i$ is the $i$-th event time and $s_i^*$ is the species that performed an action (speciated or became extinct) at time $t_i$.

The sampling probability, under the emphasis data augmentation algorithm, is

$$\log g(x|\theta) = \sum_{i\in\mathscr{M}} \left[ \log\left( \sum_{s\in\mathscr{S}_{t_i}} \lambda_{t_i,s|\theta} \right) - \log\left( N^e_{t_i^-} + 2N^o_{t_i^-} \right) + \log(\mu_0) \; - \mu_0(t^e_i - t_i) \right] - \int_{t_0}^{t_p} \left[ \sum_{s\in\mathscr{S}_r} \lambda_{r,s|\theta} \left(1 - e^{-\mu_0(t_p - r)}\right) \right] dr$$

where $N^e_{t_i^-}$ is the number of missing species just before time $t$ and $N^o_{t_i^-}$ is the number of extant species just before $t$, $\mathscr{M}$ is the set of indexes corresponding to missing speciations, and $t^e_i$ is the extinction time of the species that speciated at time $t_i$.

We are interested in the logarithm of the ratio

$$\log r(x|\theta) = \log f(x|\theta) - \log g(x|\theta).$$

In the case of constant extinction rate $\mu_{t_i,s^*_i|\theta} = \mu_0$ there are several simplifications and the log of the ratio is

$$\log r(x|\theta) = \sum_{\mathscr{H}_{spe}} \log\left(\lambda_{t_i,s^*_i|\theta}\right) - \sum_{i=1}^{p} N^o_{t_i}\mu_0(t_i - t_{i-1}) + \sum_{i\in\mathscr{M}}^{p} \left[ \log\left(N^e_{t_i^-} + 2N^o_{t_i^-}\right) - \log\left( \sum_{s\in\mathscr{S}_{t_i}} \lambda_{t_i,s|\theta} \right) \right] + \int_{t_0}^{t_p} \sum_{s\in\mathscr{S}_t} \lambda_{t,s|\theta} e^{-\mu_0(t_p - t)} dt$$

Thus, for the LDPD model we have

$$\begin{aligned} \log r(x|\theta) = & \sum_{i\in\mathscr{H}_{spe}} \log\left(\lambda_0 + \beta_N(2 - N_{t_i}) + \beta_P P'_{t_i,s^*}\right) + \\ & \sum_{i\in\mathscr{M}}^{p} \left[\log\left(N^e_{t_i} + 2N^o_{t_i}\right) - \log\left(N_{t_i}(\lambda_0 + \beta_N(2 - N_{t_i}))\right)\right] + \\ & \sum_{i=1}^{p} N^o_{t_i}\mu_0(t_i - t_{i-1}) + N_{t_i}(\lambda_0 + \beta_N(2 - N_{t_i}))\frac{e^{-\mu_0 t_p}}{\mu_0}[e^{\mu_0 t_i} - e^{\mu_0 t_{i-1}}] \end{aligned} \tag{4.4}$$

where $P'_{t,s} = \left(PDM_t - P_{t,s}\right)$. Thus, the gradients with respect to the various parameters is calculated with the partial derivatives

$$\begin{aligned} \frac{\partial \log r(x|\theta)}{\partial_{\mu_0}} = & \sum_{i=1,\ldots,p} N^o_{t_i}(t_i - t_{i-1}) + \\ & \frac{1}{\mu_0^2} N_{t_i}\left[\lambda_0 + (2 - N_{t_i})\beta_N\right]\left[e^{\mu_0(t_i - t_p)}[\mu_0(t_i - t_p) - 1] - e^{\mu_0(t_{i-1} - t_p)}[\mu_0(t_{i-1} - t_p) - 1]\right], \end{aligned} \tag{4.5}$$

$$\frac{\partial \log r(x|\theta)}{\partial_{\lambda_0}} = \sum_{i \in \mathscr{H}_{spe}} \left[ \frac{1}{\lambda_0 + \beta_N(2 - N_{t_i}) + \beta_P P'_{t_i,s^*}} \right] + \sum_{i \in \mathscr{M}_x} \left[ \frac{-N_{t_i}}{N_{t_i}(\lambda_0 + \beta_N(2 - N_{t_i}))} \right] + \sum_{i \in \{1,\dots,p\}} \left[ N_{t_i} \frac{e^{-\mu_0 t_p}}{\mu_0} \left[ e^{\mu_0 t_i} - e^{\mu_0 t_{i-1}} \right] \right], \tag{4.6}$$

$$\frac{\partial \log r(x|\theta)}{\partial_{\beta_N}} = \sum_{i \in \mathscr{H}_{spe}} \left[ \frac{(2 - N_{t_i})}{\lambda_0 + \beta_N(2 - N_{t_i}) + \beta_P P'_{t_i,s^*}} \right] + \sum_{i \in \mathscr{M}_x} \left[ \frac{-N_{t_i}}{N_{t_i}(\lambda_0 + \beta_N(2 - N_{t_i}))} \right] + \sum_{i \in \{1,\dots,p\}} \left[ N_{t_i} \frac{e^{-\mu_0 t_p}}{\mu_0} \left[ e^{\mu_0 t_i} - e^{\mu_0 t_{i-1}} \right] \right], \tag{4.7}$$

and

$$\frac{\partial \log r(x|\theta)}{\partial_{\beta_P}} = \sum_{i \in \mathscr{H}_{spe}} \left[ \frac{P'_{t_i,s^*}}{\lambda_0 + \beta_N(2 - N_{t_i}) + \beta_P P'_{t_i,s^*}} \right]$$

Thus, we have an explicit form to compute the stochastic gradient descent step and perform optimisation. The method can be used for any kind of model, but gradients need to be calculated in every case.

## 4.5. SUMMARY

Species diversification models can be used to quantify the relationship of different ecological variables with species diversification processes. Most current implemented species diversification models assume that all species are equally probable to speciate or become extinct, which does not allow to quantify the effect of the topology on the processes.

We have presented a generalised diversity-dependence model that preserves the relationship between speciation rate and species richness of previously studied models but adds an ecological advantage to species that are either more or less phylogenetically distant to the other species in the clade. With simulations we have shown that this model is flexible enough to capture a large variety of topologies. We propose this model as a standard alternative to current diversity-dependent diversification models. This model can be also complemented with the model of the previous chapter, which also takes into account dynamical carrying capacities.

Finally, we have presented an estimation method based on a stochastic gradient descent method and provide the corresponding equations to use it for the LDPD model.

# 5

# APPROXIMATING THE PROBABILITY OF CONDITIONING EVENTS IN SPECIES DIVERSIFICATION MODELS USING GENERALISED ADDITIVE MODELS

*Don't have a fixed idea in your head. Use everything you've learned until now.*

Masaaki Hatsumi

Richter F, Etienne RS, Wit EC. In preparation.

## ABSTRACT

*In cosmology, the anthropic principle describes the argument that any calculation of the probability of (intelligent) life in the universe has to take into account that the mere fact of performing such calculation presupposes intelligent life. Although not as extreme, in phylogenetic analyses similar considerations occur. For example, the probability of an observed phylogenetic tree presupposes that this tree is not empty.*

*It is common practice, therefore, in likelihood-based methods of fitting a species diversification model to an observed phylogenetic tree to condition the likelihood. Typical conditioning events include survival of the process to the present, the age of the tree, the number of species in the tree, or a combination of these. To condition a likelihood, the probability of the condition event is needed as a function of the parameters. The calculation of these probabilities is usually not trivial and often unfeasible.*

*Here, we present a general method that can be used to approximate the probability of any conditioning event under any species diversification model from which it is possible to obtain samples. This is crucial for inference in a large number of real-world scenarios. We provide an example and compare our results with probabilities that are computed using standard methods in cases in which analytic solutions are available. We find that our method is fast and accurate.*

## 5.1. INTRODUCTION

Under ideal laboratory circumstances, a probabilistic experiment will have a set of possible outcomes $y \in \mathscr{Y}$. If $\theta$ represents the parameters of the model that gives rise to the outcome $y$, then maximum likelihood estimation would aim to find $\hat{\theta} = \text{argmax}_\theta f_\theta(y)$. However, if in the real experimental setting only values within $C \subset \mathscr{Y}$ are possible, then this should be taken into consideration when performing inference. In fact, $C$ will act as a conditioning event in the estimation, i.e., $\hat{\theta} = \text{argmax}_\theta f_\theta(y|C)$.

Conditioning as part of inference of the species diversification model is important for the estimation of the speciation and extinction parameters from phylogenetic trees (Etienne *et al.*, 2016; Stadler, 2013). Several types of conditioning are commonly applied in the literature, including conditioning on crown or stem age, i.e., assuming the process started at the given age, conditioning on the survival of the clade to the present, i.e., that at least one or two species are found at the present for a tree with a stem or crown age respectively, conditioning on the number of species observed at the present, i.e., the number of tips in the phylogenetic tree, or conditioning on having at least a certain number species at the present. It has been argued that at least some conditioning is needed to remove bias in estimates, but not too much to avoiding skewing the information in the data (Etienne *et al.*, 2016; Stadler, 2013). Here, we do not further discuss which condition should be used when analysing phylogenetic data under different scenarios, but we provide a general method that can be used for any condition.

To condition a likelihood function $f_\theta(y)$ with parameters $\theta$ on a condition $C$ satisfied by our data $y \in C$, we consider the conditional likelihood

$$f_\theta(y|C) = \frac{f_\theta(y, C)}{P_\theta(C)} = \frac{f_\theta(y)}{P_\theta(C)}$$

where $P_\theta(C)$ is the probability of the condition $C$ for the parameter combination $\theta$. In this chapter, we propose a method for estimating the probability of any conditioning event by connecting our general simulation scheme for the diversification model with the theory of generalised additive models (Hastie and Tibshirani, 1990).

## 5.2. MATERIAL AND METHODS

The mathematical problem we want to solve in this chapter can be describes as follows. Assume that the species diversification is governed by a species diversification model (SDM) with rates

$$\lambda_t = g_1(x_t'\theta_1), \qquad \mu_t = g_2(x_t'\theta_2)$$

where $\lambda_t$ and $\mu_t$ are the speciation and extinction rates at time $t$, $g_i$ are arbitrarily link functions, $x_t$ is a vector of covariates or ecological variables of the diversification process at time $t$, and $\theta = (\theta_1, \theta_2)$ are the parameters that relate the speciation and extinction rates to these covariates. We want to calculate the probability that a tree with parameters $\theta$ satisfies a condition $C$.

Our framework consists of two phases: (i) the simulation or data generation phase and (ii) the GAM estimation phase, where an estimation of the probability of the desired conditioning event as a function of the parameters is calculated.

Conditioning the likelihood consists of calculating the probability that the desired condition holds for a given SDM $\mathscr{M}_\theta$ for any parameter combination. The functional relationship between $P_\theta(C)$ and $\theta$ may be complex, but we assume that $P_\theta(C)$ is a continuous function of $\theta$, in a compact subspace $\Theta_0 \subset \Theta$. We also assume that simulating the phylogenetic process according to the model $\mathscr{M}_\theta$ is possible and fast.

### 5.2.1. SIMULATION

To perform non-parametric estimation of the probability of a condition we first generate the required data in a finite grid $\Theta_G = \{\theta_1, \theta_2, ..., \theta_M\} \subset \Theta_0$ , where $M$ can be determined by the user, depending on the desired level of precision. We simulate one tree $y$ at every point of the grid. We then define the Dirac delta function $\delta_C$ that is equal to 1 at that point if the tree $y$ satisfies the desired condition $C$ and a 0 if it does not satisfy this condition, i.e.,

$$\delta_C(y) = \begin{cases} 1 & \text{if } y \in C, \\ 0 & \text{otherwise.} \end{cases}$$

The key observation that makes our framework work is that

$$P_\theta(C) = \mathbb{E}_{y\sim\theta}[\delta_C(y)]$$

Our data is given by $M$ pairs,

$$\{(\theta_1, \delta_C(y_1)), ..., (\theta_M, \delta_C(y_M))\} \tag{5.1}$$

Simulation of the phylogenetic trees is typically done in a straighforward manner using a Gillespie-type algorithm (Gillespie, 1977), which was designed in the 1970s in the context of chemical reactions. For simulations of the phylogenetic trees described in this thesis this algorithm can be used. For a detailed study on simulation methods in macroevolution see Huelsenbeck (1995).

### 5.2.2. ESTIMATION

In order to obtain the probability $P_\theta(C)$ of satisfying the condition as function of the parameters, we perform functional estimation using our simulated data. We assume that, for a continuous function $P_\theta(C)$ and a large value $K$, we can express

$$\text{logit}(P_\theta(C)) \approx \sum_{j=1}^{K} \beta_j b_j(\theta) \tag{5.2}$$

where $b_1, ..., b_K$ are basis functions. For the basis functions, we can use, for example, univariate or bivariate cubic splines (Durrleman and Simon, 1989). In Figure 5.1 we see an example of basis functions in one dimension.

Thus, given the data obtained in (5.1), we can calculate the log-likelihood of a logistic regression in the generalized additive model setting,

$$\ell^P(\beta|\theta) = \sum_{i=1}^{M} \ell_{\delta_C(y_i)}(\beta|\theta) + \text{Pen}(\beta)$$

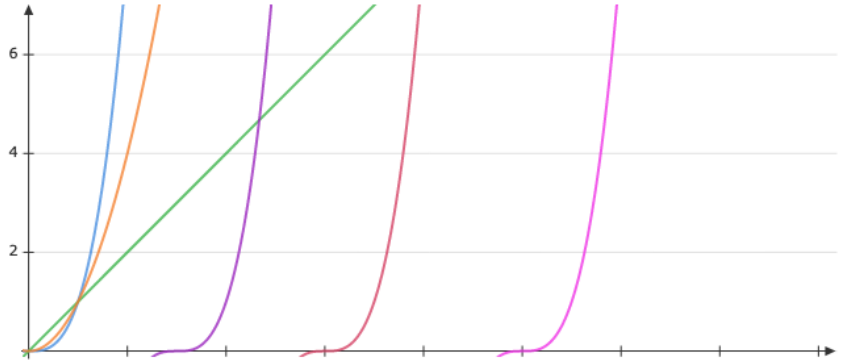



**Figure 5.1 |** Set of basis functions in one dimension, $\{t, t^2, t^3, (t-1.5)^3, , (t-3)^3, (t-5)^3\}$. The set of linear combination of them generate a large subset of continuous functions.

where Pen is typically some smoothness penalty and

$$\ell_{\delta_C(y_i)}(\beta|\theta) = \log\left(\left[\frac{g_{M,\beta}(\theta)}{1+g_{M,\beta}(\theta)}\right]^{\delta_C(y_i)}\left[\frac{1}{1+g_{M,\beta}(\theta)}\right]^{1-\delta_C(y_i)}\right)$$

and

$$g_{M,\beta}(\theta) = \sum_{j=1}^{M} \beta_j b_j(\theta).$$

Using the standard maximum likelihood estimator (Wood and Wood, 2015)

$$\hat{\beta} = \arg\max \ell^P(\beta)$$

we can predict for all $\theta \in \Theta_0$ the probability of the conditioning event as

$$P_\theta(C) = \frac{e^{g_{M,\hat{\beta}}(\theta)}}{1+e^{g_{M,\hat{\beta}}}}, \tag{5.3}$$

which is the desired probability.

Figure 5.2 contains a representation of the two phases of the method. First, we simulate trees for a grid in the parameter space, and we record the binary variable that indicates if the condition is satisfied or not. We then perform statistical inference to find the continuous function that best represents the simulated data; the estimation step is a standard generalized additive model fit (Hastie and Tibshirani, 1990). We end up with an estimation of the probability of the condition as a function of the parameters. We can make the estimation as accurate as we want by increasing the resolution of the grid.

## 5.3. APPLICATION

To check the feasibility of the method, we calculate the probability of survival of the process generated by a linear diversity-dependent diversification model LDD (Etienne

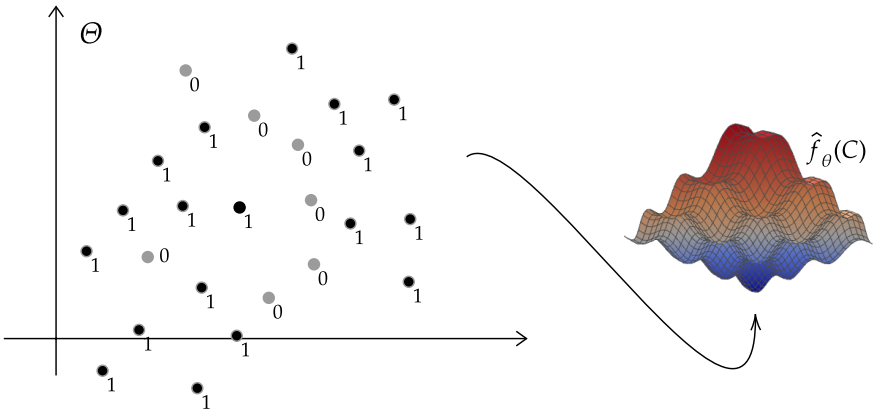



**Figure 5.2 |** Schematic representation of our framework to compute the conditional probability. Given a grid in parameter space, we simulate one tree at every point and record a zero if the tree does not satisfy the condition $C$ and a one if it does satisfy condition $C$. With the values at every point of the grid we estimate the probability surface. The estimated curve is an approximation of the desired probability function. The accuracy of this estimation is expected to increase with the resolution of the grid.

*et al.*, 2012a) introduced in the previous chapters, starting with one initial species. We assume the speciation rate to be

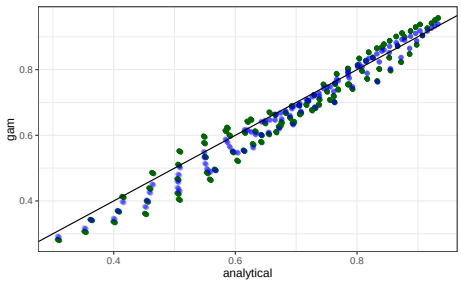
$$\lambda_{t,s} = \lambda_0 - \beta_N N_t; \qquad \lambda_0 > 0, \beta_N > 0 \tag{5.4}$$

and the extinction rate to be constant $\mu_{t,s} = \mu_0$.

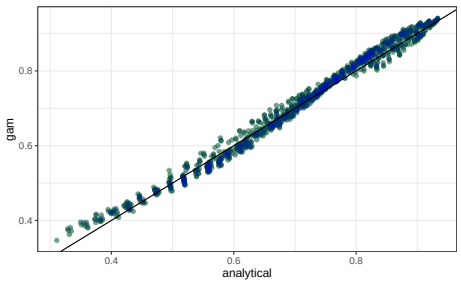
The probability of survival of the diversification process at the present can be calculated directly using the `DDD` package in `R`. We will compare this solution with the estimated solution obtained with our new method.

We approximate the function in the parameter grid where $\mu_0 \in [0.1, 0.5]$, $\lambda_0 \in [0.55, 1.5]$ and $\beta_N \in [-0.005, -0.0002]$. For this we use the `fit_gam_survival` function in the R package `emphasis`, which critically depends on the `gam` function of the `mgcv` package (Wood and Wood, 2015). When fitting a Generalised Additive Model, it is required to define the splines functions that should be used. In this application, we used univariate and bivariate splines (Nürnberger and Zeilfelder, 2000). In Figure 5.3 we plot predictions for 6 different grid resolutions in the simulation step. Each point corresponds to the probability of survival at a different parameter value; in the x-axes we compute the analytical solution of the probability and in the y-axes we plot the GAM approximation. In green points corresponds to the estimations with bivariate splines and in blue points are predictions using univariate splines, in most cases the choice of the type of spline makes little difference. We observe that estimations are quite accurate, just being a bit off in the borders of the parameter space.

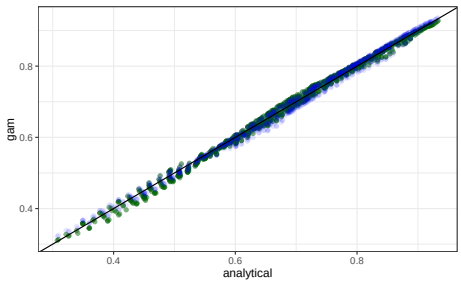
In Table 5.1, we see the computational cost (in seconds) and the error in the predictions. We calculate the error as the mean of the absolute difference between the analytical and the predicted results, across all points in the grid. We see that the error decreases



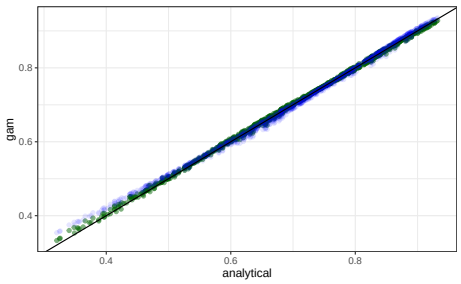

**(a)** $M = 1000$



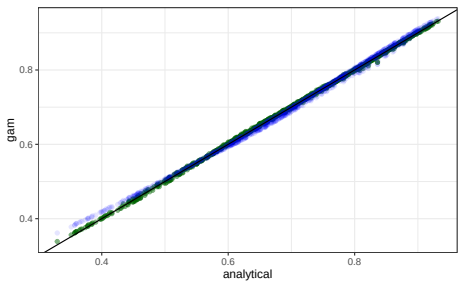

**(b)** $M = 8000$



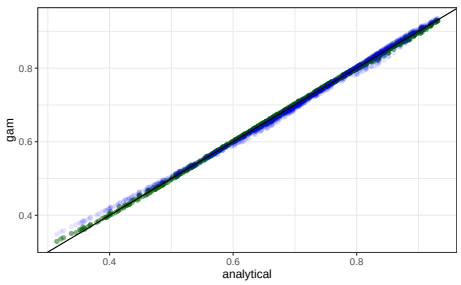

**(c)** $M = 27000$




**(d)** $M = 125000$




**(e)** $M = 421875$




**(f)** $M = 1000000$

**Figure 5.3 |** Estimation of the survival probability for different values of the number of simulations $M$. The x-axis represents the analytical results obtained with package `DDD` for various values of the parameters $(\mu_0, \lambda_0, \beta_0)$ and the y-axis corresponds to the predictions of the probability of survival with the method presented in this chapter. The black thin line represents the relationship y = x. Panels contains predictions using univariate (green) and bivariate (blue) splines for GAM estimation.

quite fast, at a small cost. $10^6$ simulations are enough to get accurate predictions.

In Table 5.2, we have a summary of the fitted models when a resolution of $100 \times 100 \times 100$ was used in an hypercube for the simulation step. We see that for the first model all parameters in the GAM are significant. We slightly prefer the bivariate model, but the difference among them is small.

With the estimated parameters, we have a linear combination of univariate splines that estimates the probability of survival of a tree governed by a LDD model. Figure 5.4

| Number of points | Simulation time [sec] | Estimation time [sec] | Error |
|---|---|---|---|
| 1000 | 0.6 | 0.3 | 0.035 |
| 8000 | 4.6 | 1.6 | 0.014 |
| 27000 | 15.5 | 4.7 | 0.008 |
| 125000 | 71.3 | 64.3 | 0.003 |
| 421875 | 237.3 | 149.3 | 0.002 |
| 1000000 | 581.8 | 446.5 | 0.002 |

**Table 5.1 |** Cost and estimation of the probability of condition with different number of simulations. Last column contains the deviance of the predicted values in relation with the analytical values. Columns 2 and 3 reports the cost (in seconds) to obtain the estimating function. The first columns contains the number of points utilised in the simulation step.

shows a visualisation of the estimated probability of survival as a function of the first two parameters.

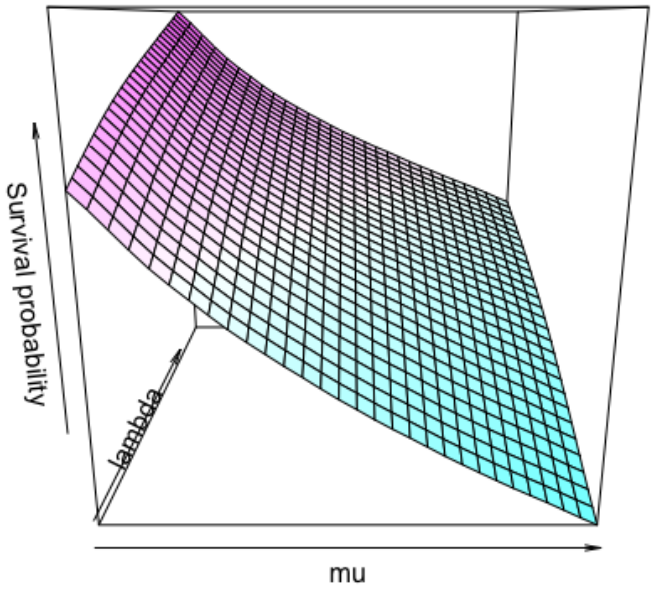



**Figure 5.4 |** Survival probability surface obtained with our new method for the LDD example.

This illustrative example shows the capacity of this method to provide accurate predictions in a few minutes. The computing times reported in table 5.1 are obtained with a standard laptop and with code that can be easily optimised (e.g. parallelised) to increase efficiency.

| | Bivariate | Univariate |
|---|---|---|
| (Intercept) | $0.94^{***}$ | $0.94^{***}$ |
| | (0.00) | (0.00) |
| EDF: s($\mu_0$,$\lambda_0$) | $13.16^{***}$ | |
| | (17.30) | |
| EDF: s($\mu_0$,$\beta_N$) | $10.59^{***}$ | |
| | (13.87) | |
| EDF: s($\lambda_0$,$\beta_N$) | $0.48^{***}$ | |
| | (27.00) | |
| EDF: s($\mu_0$) | | $6.25^{***}$ |
| | | (7.41) |
| EDF: s($\beta_N$) | | 2.73 |
| | | (3.40) |
| EDF: s($\lambda_0$) | | $4.12^{***}$ |
| | | (5.08) |
| AIC | 1132386.62 | 1132701.08 |
| BIC | 1132684.69 | 1132867.76 |
| Log Likelihood | −566168.09 | −566336.44 |
| Deviance | 1132336.17 | 1132672.87 |
| Deviance explained | 0.07 | 0.07 |
| Dispersion | 1.00 | 1.00 |
| $R^2$ | 0.09 | 0.09 |
| GCV score | 0.13 | 0.13 |
| Num. obs. | 1000000 | 1000000 |
| Num. smooth terms | 3 | 3 |

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

**Table 5.2 |** Summary of estimated GAMs using univariate and bivariate basis splines.

## 5.4. DISCUSSION

We have presented a methodology to estimate the probability of any condition or outcome that can be simulated, given an arbitrary species diversification model. This represents an application of the theory of generalised additive models in the context of networks and trees.

As a proof of concept, we calculate the probability of survival of any species after a fixed time for the species diversification process under a LDD model. We provide a functional polynomial form that represents an estimation of the survival probability of the process. We found that the method performs well at a low computational cost. However, it is only tested so far in a three-parameter model. We did not explore how the efficiency of the method decreases with an increase in parameters. For high dimensional problems, where the diversification models depends on many covariates, adaptations to high-dimensional GAM modelling is possible (Meier *et al.*, 2009, e.g.).

We add two remarks

- The limits of the grid are important. In the example presented here predictions can be used in the cube $\mu_0 \in [0.1, 0.5]$, $\lambda_0 \in [0.55, 1.5]$ and $\beta_N \in [-0.005, -0.0002]$ only. Extrapolations must be avoided.
- The estimation of the function is based on a smooth polynomial (splines) approximation, so to have proper approximation we assume that the probability of the condition is a smooth function, which is biologically sensible to assume.

In conclusion, we recommend using this method when conditioning a species diversification process and the probability of the condition is not available, but simulation of the process is possible. This method is complementary with the theory presented throughout this thesis. In Chapters 2, 3 and 4 we have presented a likelihood approach in different contexts for statistical inference, however, to keep the focus on the statistical methodologies presented we did not consider any conditioning to the estimated likelihoods. The estimated probability can be incorporated in both the MCEM algorithm presented in Chapters 2 and 3 and in the SGD algorithm presented in Chapter 4. In all cases the condition probability have to be computed only once at the start of the statistical inference routine, unless the MCEM or the SGD goes to parts of parameter space that weren't explored by the GAM; then a new simulation needs to be done including the new explored parameter space.

# 6

# FURTHER CONSIDERATIONS REGARDING SPECIES DIVERSIFICATION MODELLING

*The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously parallel.*

Darwin C. 1879

The intersection between macroevolution and statistical modelling represents a fundamental and growing area of research for the understanding of how species diversified. In this thesis, using combinations of statistical methods, I have presented methodological tools that will contribute to the study of species diversification in a rather general way, i.e., applying to a wide variety of scenarios in macroecology and macroevolution. Still, despite the potential of the methods presented here, we have focused all the applications on a particular class of models which considers diversity as the primary driver of diversification. Diversity dependent diversification models possess attractive properties in both evolutionary biology and mathematics.

In evolutionary biology, the incorporation of diversity-dependent diversification models is a sensible, quantitative way to test the influence of ecological limits on diversification. Until now, diversity-dependent diversification models have used species richness as a proxy for diversity. This is a substantial simplification considering that different species may contribute differently to ecological limits; by considering only species richness we assume that all species in a clade compete in the same way for the same niches. Throughout this thesis, I have incorporated phylogenetic diversity, generalising diversity-dependent diversification models, the inference of which has so far not been possible with current statistical methods. With the incorporation of phylogenetic diversity we take into account the genetic difference among species. Phylodiversity in combination with species richness includes dynamic carrying capacities instead of fixed ones. When considering

pairwise phylogenetic diversity, we also consider variable ecological interactions among species.

Mathematically, both species richness and phylogenetic diversity are relevant properties of phylogenetic trees. Moreover, given that this information is provided by the tree itself, there is no need for imputing other unobserved variables than the extinct species. This is relevant for the accuracy of the methods. Its simplicity provides an elegant way to deal and incorporate other factors, such as ecological similarities between lineages. As William of Ockham suggested, a mathematical model should aim for capturing the behaviour of a complex system with a maximum level of simplicity (Schaffer, 2015). Our generalisations to diversity dependent diversification models share these properties.

The applications presented here represent a contribution to statistical network sciences in biology. We do not only demonstrate the feasibility of optimisation methods such as the EM and the SGD algorithms in point processes describing trees, but we also provide insights into the design of efficient data augmentation algorithms for trees and networks.

The journey of this thesis has met with a lot of trial and error, exploring several approaches that end up not being as appropriate as the MCEM and SDG methods described in Chapters 2, 3 and 4. In practice what I have presented in this thesis is a small fraction of what I have tried in order to provide an efficient way to perform statistical inference on species diversification processes. Moreover, I am aware that this is a small contribution and a first step into the long-term development of a general theory that will eventually improve the methods here presented. In the next sections, I discuss the limitations of these methods and possible directions for improvement and development.

## 6.1. LIMITATIONS IN SYSTEMATIC BIOLOGY AND DIRECTIONS FOR IMPROVEMENT

### 6.1.1. INCOMPLETE SAMPLING AND DIFFERENT LEVELS OF ORGANISMS

Throughout this thesis, I have assumed that the phylogenies contain all the extant species of the biological system of interest. In practice, that is seldom true. Even though phylogenies are becoming every year more complete and accurate, incomplete sampling is the most probable situation for most of the phylogenies, especially in groups such as insects or non-vertebrates, to name two examples. Assuming that the sampled phylogeny is complete is a common practice, but further research should consider the consequences of incomplete sampling in phylogenetic analysis. Some authors have considered incomplete sampling in their methodologies (Carstens and Knowles, 2007; Wiuf, 2018), but there is still a long way to go.

For the methodologies here presented, the natural extension to consider incomplete sampling would be to slightly modify the data augmentation scheme proposed on section 3.3.3, allowing a less restrictive space of trees, where the sampled full trees do not necessarily have the same number of species at present. The theory and implementation of this extension are easy, but assumptions need to be made. For instance, some current methods provide the option to add "the number of missing species" at present, this is a

very strong assumption. A more relaxed assumption would be to assume a probability distribution to the species sampling scheme. Such an assumption is also possible to include in our frameworks in a nearly straightforward mathematically and computationally way; in contrast, biologically, the assumed probability distribution would always be debatable.

### 6.1.2. EXTINCTION DYNAMICS

In order to compare and analyse the generalisations presented here, I decided to focus in all models on dynamic speciation rates while the extinction rates are assumed to be constant. However, the methodologies presented in this thesis are more general and in principle can deal with non-constant extinction rates. In all illustrations and experiments, the same methods can be used for non-constant extinction rates with almost no additional work. I suggest that further research incorporates this characteristic when testing the hypothesis of non-constant extinction rates. A natural generalisation that align with this thesis would be the diversity-dependent extinction models, supposing a linear extinction rate as a function of diversity (i.e species richness and phylogenetic diversity).

### 6.1.3. IMPLEMENTING THE GENERAL CLASS OF MODELS

The methods presented here have been developed to answer the question: "What factors can play a role in species diversification?" As mentioned in the introduction, despite the great potential of the general class of models that our new inference methods accommodates, I did all the illustrations in diversity-dependence models or generalisations of them. In future work, depending on the focus of the different required analysis, incorporating extreme events, climate or other time-dependent functions, individual characteristics of species or other factors could be incorporated.

In Figure 6.1 I illustrate a process where each species has a binary trait represented by a circle or a triangle. This can be, for instance, presence/absence of legs in squamates or viviparity. Given that the species-level covariate data is typically only available at the present, as shown in Figure 6.1a, there are a large number of compatible covariate histories over which any inference procedure should integrate. In principle, our *emphasis* simulation and inference framework is capable of dealing with such situations, but it is not clear to what extent the methods presented here can handle a large number of species-level covariates and it is expected that if the unavailable data on these covariates is large, the integration across them will be challenging.

## 6.2. DIRECTIONS FOR STATISTICAL METHODS

I have presented a number of statistical methods, which I would classify in three categories: statistical network processes modelling, data augmentation algorithms and parametric statistical inference. These methodologies open up an endless set of combinations.

In the statistical modelling I consider the theory of point process, assuming that speciation and extinction can be realistically generated by combinations of non-homogeneous Poisson processes. That theory was primarily developed by Yule in the 1920s, Kendall

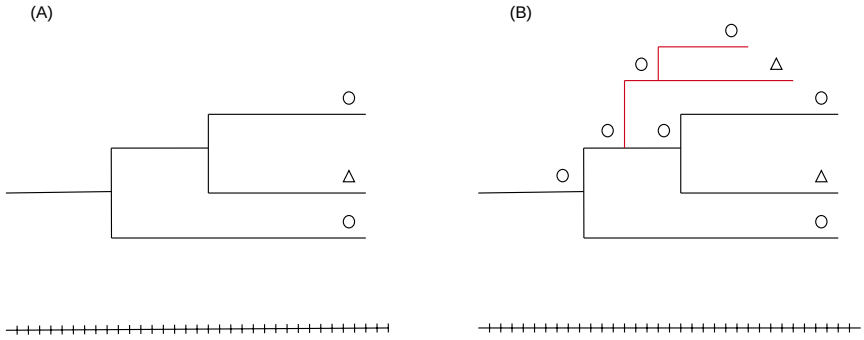



**Figure 6.1 |** a) Extant phylogenetic tree without extinctions and b) complete phylogentic tree with extinctions. Both trees are shown with a binary trait indicators.

in the 1940s and Nee in the 1990s. All this work and subsequent developments have solved a great variety of problems, but there has never been any attempt to define a general class of species diversification models, as I have tried to present in this thesis. One of the main reasons why inference in this class of models has remained elusive is the complexity involved in the underlying system of stochastic differential equations given by the combination of point processes involved in the macroevolutionary dynamics. I have provided an alternative to direct likelihood calculations, by means of an importance sampling simulation scheme. In principle, this may allow to integrate inference of a general class of species diversification models in one single framework. Still, a lot of work is needed; for instance, in the NHPP I do not allow multiple speciations at the same time or protracted speciation, i.e., speciation events that take time (i.e. not instantaneous).

For statistical inference I have proposed two alternatives to calculate and optimise the likelihood of the species diversification process under incomplete information. One is the MCEM algorithm developed in Chapters 2 and 3. The other is the SGD method developed in Chapter 4. These approaches are two examples of likelihood methods combined with data augmentation through simulations. Both are methods to optimise the likelihood and find the maximum likelihood estimator for complex processes where the likelihood of the observed process is impossible to be calculated directly, but for which the augmented process likelihood is much easier. Other approaches, such as Bayesian approaches or alternative optimisation algorithms such as the SAEM algorithm or its variations, have not been fully explored in this thesis.

Data augmentation algorithms (DAAs) are powerful statistical tools for studying the full or augmented process likelihood, as they provide a solution in cases where the likelihood for the original data is difficult or impossible to calculate, such as is the case in general species diversification processes where only the reconstructed tree is available. In this thesis I have provided two DAAs: (i) a uniform sampler that augment trees independent of the model parameters by simulating branching times and topology uniformly and

(ii) an efficient importance sampler that considers the parameters of the diversification model in order to sample trees in close accordance to the generative process. Although I have implemented the DAA inference methods in an R package, computational efficiency was not the main focus of this thesis and I believe that this can be improved as well.

## 6.3. EVOLUTIONARY TREES APPLICATIONS, BEYOND BIOLOGY

The theory presented in this thesis describes a *theory of diversification* in a general sense. Thus, nothing stops us from applying this framework in contexts different from evolutionary biology where also diversification processes take place. One can think of language evolution (Greenhill *et al.*, 2010; Whitfield, 2008; Zhang *et al.*, 2020) or cultural evolution (Creanza *et al.*, 2017), to name just two examples. More abstractly, tree-like diversification happens in many different processes. In Figure 6.2 I show nine tree-shape phenomena taken from many different fields. This is a small sample of tree-like diversification in nature.

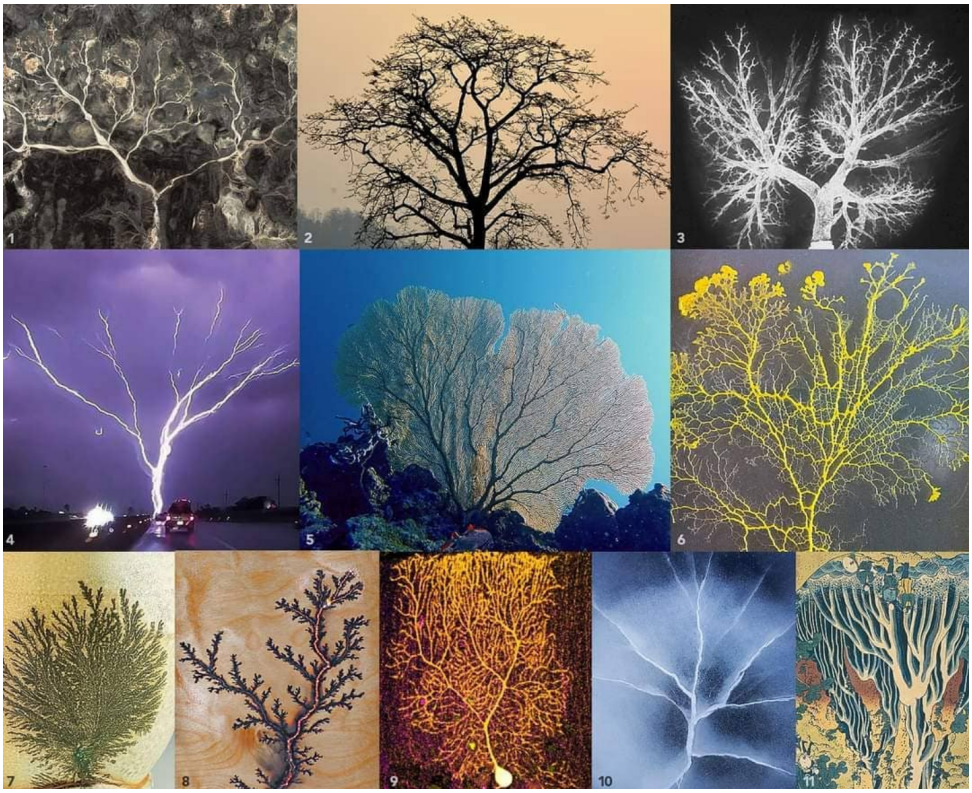



**Figure 6.2 |** 1. Image of river from space. 2. Tree. 3. Human bronchus (upside down). 4. Upward lightning. 5. Coral. 6. Slime mold. 7. Mocha diffusion. 8. Lichtenberg figures on wood. 9. Human neuron. 10. Cracked ice. 11. Waterfall - Katsushika Hokusai (1831) (upside down)

Moreover, even within biology, multiple other applications can benefit from the statistical methodologies here presented. In this thesis, we have focused on species-level trees. However, diversification processes happen at all levels of organisms and scales of times (Aldous *et al.*, 2008; Stadler and Bokma, 2013).

## 6.4. NETWORK SCIENCES APPLICATIONS, BEYOND TREES

Trees are the most common representation for the diversification of species. However, other mathematical objects have been suggested to describe species diversification, such as phylogenetic networks, phylogenetic cactus or phylogenetic corals (Ragan, 2009; Podani, 2017), among others. Statistical network science is a growing area of research (Molontay and Nagy, 2019) and it has great potential to contribute to the field of evolutionary phylogenetics (Huson and Bryant, 2006; Chamberlain *et al.*, 2014; Kunin *et al.*, 2005; Bandelt, 1995). Moreover, biological networks can also be potential drivers of evolutionary processes and thus incorporated as covariates in species diversificaion models (Farajtabar *et al.*, 2017).

In Figure 6.3, I show an example of a phylogenetic network where different biological process are incorporated, generalising a phylogenetic tree (Schliep *et al.*, 2016). I am convinved that all methods presented in this thesis can be generalised to networks in a mathematically natural way. Future research should consider this direction as another generalisation for species diversification models in order to describe macroevolutionary processes more realistically.

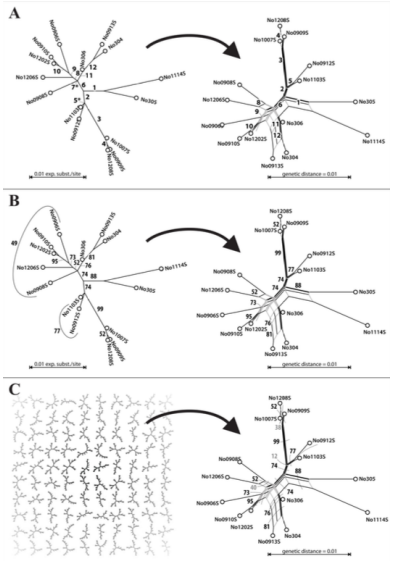



**Figure 6.3 |** Phylogenetic network from Schliep *et al.* (2016)

# REFERENCES

M. A. Ragan, *Trees and networks before and after darwin,* Biology direct **4**, 43 (2009).

W. Jetz, G. H. Thomas, J. B. Joy, K. Hartmann, and A. O. Mooers, *The global diversity of birds in space and time,* Nature **491**, 444 (2012).

N. S. Upham, J. A. Esselstyn, and W. Jetz, *Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation,* PLoS biology **17**, e3000494 (2019).

S. Ramírez-Barahona, H. Sauquet, and S. Magallón, *The delayed and geographically heterogeneous diversification of flowering plant families,* Nature Ecology & Evolution **4**, 1232 (2020).

L. A. Hug, B. J. Baker, K. Anantharaman, C. T. Brown, A. J. Probst, C. J. Castelle, C. N. Butterfield, A. W. Hernsdorf, Y. Amano, K. Ise, Y. Suzuki, N. Dudek, D. A. Relman, K. M. Finstad, R. Amundson, B. C. Thomas, and J. F. Banfield, *A new view of the tree of life,* Nature Microbiology **1**, 16048 (2016).

S. B. Hedges, J. Marin, M. Suleski, M. Paymer, and S. Kumar, *Tree of life reveals clock-like speciation and diversification,* Molecular Biology and Evolution **32**, 835 (2015a), arXiv:1412.4312v1 .

S. Patel, R. T. Kimball, and E. L. Braun, *Error in Phylogenetic Estimation for Bushes in the Tree of Life,* **1**, 1 (2013).

M. Pagel, *Inferring the historical patterns of biological evolution,* Nature **401**, 877 (1999).

D. J. Aldous, *Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today,* Statist. Sci **16**, 23 (2001).

D. G. Kendall *et al.*, *On the generalized" birth-and-death" process,* The annals of mathematical statistics **19**, 1 (1948).

S. J. Gould, D. M. Raup, J. J. Sepkoski Jr, T. J. Schopf, and D. S. Simberloff, *The shape of evolution: a comparison of real and random clades,* Paleobiology , 23 (1977).

S. M. Stanley, *Effects of competition on rates of evolution, with special reference to bivalve mollusks and mammals,* Systematic Zoology **22**, 486 (1973).

D. M. Raup, S. J. Gould, T. J. M. Schopf, and D. S. Simberloff, *Stochastic models of phylogeny and the evolution of diversity,* The Journal of Geology **81**, 525 (1973).

J. F. Reynolds, *ON ESTIMATING THE PARAMETERS OF A BIRTH-DEATH PROCESS,* Australian & New Zealand Journal of Statistics **15**, 35 (1973).

S. Nee, R. M. May, and P. H. Harvey, *The reconstructed evolutionary process,* Philosophical Transactions of the Royal Society of London B: Biological Sciences **344**, 305 (1994).

M. G. B. Blum and O. François, *Which random processes describe the tree of life? A large-scale study of phylogenetic tree imbalance,* Systematic Biology **55**, 685 (2006).

F. S. Caron and M. R. Pie, *The phylogenetic signal of diversification rates,* Journal of Zoological Systematics and Evolutionary Research , 1 (2020).

L. Popovic, *Asymptotic genealogy of a critical branching process,* The Annals of Applied Probability **14**, 2120 (2004).

O. Hagen, K. Hartmann, M. Steel, and T. Stadler, *Age-dependent speciation can explain the shape of empirical phylogenies,* Systematic biology **64**, 432 (2015).

P. Descombes, T. Gaboriau, C. Albouy, C. Heine, F. Leprieur, and L. Pellissier, *Linking species diversification to palaeo-environmental changes: A process-based modelling approach,* Global Ecology and Biogeography **27**, 233 (2018).

E. E. Goldberg, L. T. Lancaster, and R. H. Ree, *Phylogenetic inference of reciprocal effects between geographic range evolution and diversification,* Systematic Biology **60**, 451 (2011).

D. Silvestro, J. Schnitzler, and G. Zizka, *A Bayesian framework to estimate diversification rates and their variation through time and space,* BMC evolutionary biology **11**, 311 (2011).

S. Höhna, *The time-dependent reconstructed evolutionary process with a key-role for mass-extinction events,* Journal of theoretical biology **380**, 321 (2015), arXiv:1312.2392 .

W. P. Maddison, P. E. Midford, and S. P. Otto, *Estimating a binary character's effect on speciation and extinction,* Systematic biology **56**, 701 (2007).

J. M. Beaulieu and B. C. O'Meara, *Detecting hidden diversification shifts in models of trait-dependent speciation and extinction,* Systematic biology **65**, 583 (2016).

L. Herrera-Alsina, P. van Els, and R. S. Etienne, *Detecting the dependence of diversification on multiple traits from phylogenetic trees and trait data,* Systematic biology **68**, 317 (2019).

D. A. Rasmussen and T. Stadler, *Coupling adaptive molecular evolution to phylodynamics using fitness-dependent birth-death models,* eLife **8**, e45562 (2019).

D. Schluter, *Speciation, Ecological Opportunity, and Latitude: (American Society of Naturalists Address),* The American Naturalist **187**, 1 (2016).

F. L. Condamine, J. Rolland, and H. Morlon, *Assessing the causes of diversification slowdowns: temperature-dependent and diversity-dependent models receive equivalent support,* Ecology letters **22**, 1900 (2019).

R. S. Etienne, B. Haegeman, T. Stadler, T. Aze, P. N. Pearson, A. Purvis, and A. B. Phillimore, *Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record,* Proceedings of the Royal Society B: Biological Sciences **279**, 1300 (2012a).

A. Lambert, H. Morlon, and R. S. Etienne, *The reconstructed tree in the lineage-based model of protracted speciation,* Journal of mathematical biology **70**, 367 (2015).

R. S. Etienne, H. Morlon, and A. Lambert, *Estimating the duration of speciation from phylogenies,* Evolution **68**, 2430 (2014).

D. L. Rabosky, *Extinction rates should not be estimated from molecular phylogenies,* Evolution: International Journal of Organic Evolution **64**, 1816 (2010).

D. L. Rabosky, *Challenges in the estimation of extinction from molecular phylogenies: a response to beaulieu and o'meara,* Evolution **70**, 218 (2016).

L. J. Harmon and S. Harrison, *Species diversity is dynamic and unbounded at local and continental scales,* The American Naturalist **185**, 584 (2015).

D. L. Rabosky and A. H. Hurlbert, *Species richness at continental scales is dominated by ecological limits,* The American Naturalist **185**, 572 (2015).

R. S. Etienne, B. Haegeman, T. Stadler, T. Aze, P. N. Pearson, A. Purvis, and A. B. Phillimore, *Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record,* Proc. R. Soc. B , rspb20111439 (2012b).

C. R. Marshall and T. B. Quental, *The uncertain role of diversity dependence in species diversification and the need to incorporate time-varying carrying capacities,* Philosophical Transactions of the Royal Society B: Biological Sciences **371**, 20150217 (2016).

A. E. Magurran, *Measuring biological diversity* (John Wiley & Sons, 2013).

A. Chao, C. H. Chiu, and L. Jost, *Unifying species diversity, phylogenetic diversity, functional diversity, and related similarity and differentiation measures through hill numbers,* Annual Review of Ecology, Evolution, and Systematics **45**, 297 (2014).

G. Laudanno, B. Haegeman, D. L. Rabosky, and R. S. Etienne, *Detecting Lineage-Specific Shifts in Diversification: A Proper Likelihood Approach,* Systematic Biology **0**, 1 (2020a).

S. B. Heard, *Patterns in phylogenetic tree balance with variable and evolving speciation rates,* Evolution **50**, 2141 (1996).

H. M. Savage, *The shape of evolution: systematic tree topology,* Biological Journal of the Linnean Society **20**, 225 (1983).

A. Mir, F. Rosselló, and Others, *A new balance index for phylogenetic trees,* Mathematical Biosciences **241**, 125 (2013).

G. Casella and R. L. Berger, *Statistical inference*, Vol. 2 (Duxbury Pacific Grove, CA, 2002).

J. Pfanzagl, *Parametric statistical theory* (Walter de Gruyter, 2011).

R. Carmona and F. Delarue, *The master equation for large population equilibriums,* in *Stochastic analysis and applications 2014* (Springer, 2014) pp. 77–128.

L. Xu and M. I. Jordan, *On convergence properties of the EM algorithm for Gaussian mixtures,* Neural computation **8**, 129 (1996).

T. Gernhard, *The conditioned reconstructed process,* Journal of theoretical biology **253**, 769 (2008).

R. S. Etienne, A. L. Pigot, and A. B. Phillimore, *How reliably can we infer diversity-dependent diversification from phylogenies?* Methods in Ecology and Evolution **7**, 1092 (2016).

T. Stadler, *How can we improve accuracy of macroevolutionary rate estimates?* Systematic Biology **62**, 321 (2013).

E. Wit, E. v. d. Heuvel, and J.-W. Romeijn, *'all models are wrong...': an introduction to model uncertainty,* Statistica Neerlandica **66**, 217 (2012).

T. Janzen, S. Höhna, and R. S. Etienne, *Approximate bayesian computation of diversification rates from molecular phylogenies: introducing a new efficient summary statistic, the nltt,* Methods in Ecology and Evolution **6**, 566 (2015).

P. Lemey, M. Salemi, and A.-M. Vandamme, *The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing* (Cambridge University Press, 2009).

H. V. Cornell, *Is regional species diversity bounded or unbounded?* Biological Reviews **88**, 140 (2013).

T. H. Ezard, T. Aze, P. N. Pearson, and A. Purvis, *Interplay between changing climate and species' ecology drives macroevolutionary dynamics,* Science **332**, 349 (2011).

T. G. Barraclough, *How do species interactions affect evolutionary dynamics across whole communities?* Annual Review of Ecology, Evolution, and Systematics **46**, 25 (2015).

E. Lewitus and H. Morlon, *Detecting environment-dependent diversification from phylogenies: a simulation study and some empirical illustrations,* Systematic biology **67**, 576 (2017).

G. G. Mittelbach, D. W. Schemske, H. V. Cornell, A. P. Allen, J. M. Brown, M. B. Bush, S. P. Harrison, A. H. Hurlbert, N. Knowlton, H. A. Lessios, *et al.*, *Evolution and the latitudinal diversity gradient: speciation, extinction and biogeography,* Ecology letters **10**, 315 (2007).

V. J. Lynch, *Live-birth in vipers (viperidae) is a key innovation and adaptation to global cooling during the cenozoic,* Evolution: International Journal of Organic Evolution **63**, 2457 (2009).

E. Paradis, *Statistical analysis of diversification with species traits,* Evolution **59**, 1 (2005).

R. G. FitzJohn, W. P. Maddison, and S. P. Otto, *Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies,* Systematic biology **58**, 595 (2009).

H. Morlon, *Phylogenetic approaches for studying diversification,* Ecology letters **17**, 508 (2014).

R. E. Ricklefs, *Estimating diversification rates from phylogenetic information,* Trends in ecology & evolution **22**, 601 (2007).

T. Stadler, *Inferring speciation and extinction processes from extant species data,* Proceedings of the National Academy of Sciences **108**, 16145 (2011).

S. Höhna, T. Stadler, F. Ronquist, and T. Britton, *Inferring speciation and extinction rates under different sampling schemes,* Molecular biology and evolution **28**, 2577 (2011).

R. F. Serfozo, *Point processes,* Handbooks in operations research and management science **2**, 1 (1990).

D. J. Daley and D. Vere-Jones, *An introduction to the theory of point processes: volume II: general theory and structure* (Springer Science & Business Media, 2007).

D. T. Gillespie, *A general method for numerically simulating the stochastic time evolution of coupled chemical reactions,* Journal of computational physics **22**, 403 (1976).

D. T. Gillespie, *Exact stochastic simulation of coupled chemical reactions,* The journal of physical chemistry **81**, 2340 (1977).

H. C. Tijms, *Stochastic models: an algorithmic approach*, Vol. 994 (John Wiley & Sons Chichester, 1994).

J. P. Castillo, M. Verdú, and A. Valiente-Banuet, *Neighborhood phylodiversity affects plant performance,* Ecology **91**, 3656 (2010).

A. J. Dobson and A. Barnett, *An introduction to generalized linear models* (CRC press, 2008).

D. L. Rabosky and I. J. Lovette, *Explosive evolutionary radiations: decreasing speciation or increasing extinction through time?* Evolution **62**, 1866 (2008).

R. P. Freckleton, A. B. Phillimore, and M. Pagel, *Relating traits to diversification: a simple test,* The American Naturalist **172**, 102 (2008).

S. Hoehna, W. A. Freyman, Z. Nolen, J. Huelsenbeck, M. R. May, and B. R. Moore, *A Bayesian Approach for Estimating Branch-Specific Speciation and Extinction Rates,* bioRxiv , 555805 (2019).

F. Wrenn, *General birth-death processes: probabilities, inference, and applications,* (2012).

A. Gavryushkin and A. J. Drummond, *The space of ultrametric phylogenetic trees,* Journal of theoretical biology **403**, 197 (2016).

A. Gavryushkin, C. Whidden, and F. Matsen, *The combinatorics of discrete time-trees: theory and open problems,* bioRxiv , 063362 (2016).

A. A. P. Dempster, N. M. N. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm,* Journal of the royal statistical society. Series B (methodological) **39**, 1 (1977), arXiv:0710.5696v2 .

G. C. Wei and M. A. Tanner, *A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms,* Journal of the American statistical Association **85**, 699 (1990).

G. McLachlan and T. Krishnan, *The EM algorithm and extensions*, Vol. 382 (John Wiley & Sons, 2007).

Z. Kasa, *Generating and ranking of dyck words,* arXiv preprint arXiv:1002.2625 (2010).

A. K. Zvonkin, *Enumeration of weighted plane trees,* arXiv preprint arXiv:1404.4836 (2014).

G. Laudanno, B. Haegeman, and R. S. Etienne, *Additional analytical support for a new method to compute the likelihood of diversification models,* bioRxiv **82**, 693176 (2019).

K. A. Jønsson, P.-H. Fabre, S. A. Fritz, R. S. Etienne, R. E. Ricklefs, T. B. Jørgensen, J. Fjeldså, C. Rahbek, P. G. Ericson, F. Woog, *et al.*, *Ecological and evolutionary determinants for the adaptive radiation of the madagascan vangas,* Proceedings of the National Academy of Sciences **109**, 6620 (2012).

D. P. Faith, *Conservation evaluation and phylogenetic diversity,* Biological conservation **61**, 1 (1992).

B. Delyon, M. Lavielle, E. Moulines, *et al.*, *Convergence of a stochastic approximation version of the em algorithm,* The Annals of Statistics **27**, 94 (1999).

S. Richardson and P. J. Green, *On bayesian analysis of mixtures with an unknown number of components (with discussion),* Journal of the Royal Statistical Society: series B (statistical methodology) **59**, 731 (1997).

S. J. Greenhill, Q. D. Atkinson, A. Meade, and R. D. Gray, *The shape and tempo of language evolution,* Proceedings of the Royal Society B: Biological Sciences **277**, 2443 (2010).

R. Mace and C. J. Holden, *A phylogenetic approach to cultural evolution,* Trends in ecology & evolution **20**, 116 (2005).

T. D. Walker and J. W. Valentine, *Equilibrium models of evolutionary species diversity and the number of empty niches,* The American Naturalist **124**, 887 (1984).

D. L. Rabosky, *Ecological limits on clade diversification in higher taxa,* The American Naturalist **173**, 662 (2009).

F. L. Condamine, *Limited by the roof of the world: mountain radiations of apollo swallowtails controlled by diversity-dependence processes,* Biology letters **14**, 20170622 (2018).

G. C. Gibb, F. L. Condamine, M. Kuch, J. Enk, N. Moraes-Barros, M. Superina, H. N. Poinar, and F. Delsuc, *Shotgun mitogenomics provides a reference phylogenetic framework and timescale for living xenarthrans,* Molecular Biology and evolution **33**, 621 (2016).

R. L. Cunha, C. Patrão, and R. Castilho, *Different diversity-dependent declines in speciation rate unbalances species richness in terrestrial slugs,* Scientific reports **7**, 16198 (2017).

C. Pouchon, A. Fernández, J. M. Nassar, F. Boyer, S. Aubert, S. Lavergne, and J. Mavárez, *Phylogenomic analysis of the explosive adaptive radiation of the espeletia complex (asteraceae) in the tropical andes,* Systematic Biology **67**, 1041 (2018).

X. Chen, A. R. Lemmon, E. M. Lemmon, R. A. Pyron, and F. T. Burbrink, *Using phylogenomics to understand the link between biogeographic origins and regional diversification in ratsnakes,* Molecular phylogenetics and evolution **111**, 206 (2017).

J. N. Pinto-Ledezma, L. M. Simon, J. A. F. Diniz-Filho, and F. Villalobos, *The geographical diversification of furnariides: the role of forest versus open habitats in driving species richness gradients,* Journal of biogeography **44**, 1683 (2017).

J. A. McGuire, C. C. Witt, J. Remsen Jr, A. Corl, D. L. Rabosky, D. L. Altshuler, and R. Dudley, *Molecular phylogenetics and the diversification of hummingbirds,* Current Biology **24**, 910 (2014).

R. A. Pyron and J. J. Wiens, *Large-scale phylogenetic analyses reveal the causes of high tropical amphibian diversity,* Proceedings of the Royal Society B: Biological Sciences **280**, 20131622 (2013).

L. Xu and R. S. Etienne, *Detecting local diversity-dependence in diversification,* Evolution **72**, 1294 (2018).

L. H. Liow, T. B. Quental, and C. R. Marshall, *When can decreasing diversification rates be detected with molecular phylogenies and the fossil record?* Systematic Biology **59**, 646 (2010).

L. Herrera-Alsina, A. L. Pigot, H. Hildenbrandt, and R. S. Etienne, *The influence of ecological and geographic limits on the evolution of species distributions and diversity,* Evolution **72**, 1978 (2018).

M. M. Kling, B. D. Mishler, A. H. Thornhill, B. G. Baldwin, and D. D. Ackerly, *Facets of phylodiversity: evolutionary diversification, divergence and survival as conservation targets,* Philosophical Transactions of the Royal Society B **374**, 20170397 (2018).

S. Scheiner, E. Kosman, S. Presley, and M. Willig, *The components of biodiversity, with a particular focus on phylogenetic information,* Ecology and Evolution **7**, 6444 (2017).

T. Laity, S. W. Laffan, C. E. González-Orozco, D. P. Faith, D. F. Rosauer, M. Byrne, J. T. Miller, D. Crayn, C. Costion, C. C. Moritz, *et al.*, *Phylodiversity to inform conservation policy: An australian example,* Science of the Total Environment **534**, 131 (2015).

D. P. Faith and A. M. Baker, *Phylogenetic diversity (pd) and biodiversity conservation: some bioinformatics challenges,* Evolutionary bioinformatics **2**, 117693430600200007 (2006).

J. Cantalapiedra, T. Aze, M. Cadotte, G. Dalla Riva, D. Huang, F. Mazel, M. Pennell, M. Ríos, and A. Mooers, *Conserving evolutionary history does not result in greater diversity over geological time scales,* Proceedings of the Royal Society B **286**, 20182896 (2019).

F. Mazel, M. W. Pennell, M. W. Cadotte, S. Diaz, G. V. Dalla Riva, R. Grenyer, F. Leprieur, A. O. Mooers, D. Mouillot, C. M. Tucker, *et al.*, *Prioritizing phylogenetic diversity captures functional diversity unreliably,* Nature communications **9**, 2888 (2018).

J. Stadler, S. Klotz, R. Brandl, and S. Knapp, *Species richness and phylogenetic structure in plant communities: 20 years of succession,* Web Ecology **17**, 37 (2017).

C. Tucker, M. Cadotte, S. Carvalho, J. Davies, S. Ferrier, S. Fritz, R. Grenyer, M. Helmus, L. Jin, A. Mooers, S. Pavoine, O. Purschke, D. Redding, D. Rosauer, M. Winter, and F. Mazel, *A guide to phylogenetic metrics for conservation, community ecology and macroecology,* Biological Reviews **92**, n/a (2016).

C. O. Webb, G. S. Gilbert, and M. J. Donoghue, *Phylodiversity-dependent seedling mortality, size structure, and disease in a bornean rain forest,* Ecology **87**, S123 (2006).

C. Violle, D. R. Nemergut, Z. Pu, and L. Jiang, *Phylogenetic limiting similarity and competitive exclusion,* Ecology letters **14**, 782 (2011).

G. C. Costa, D. O. Mesquita, G. R. Colli, and L. J. Vitt, *Niche expansion and the niche variation hypothesis: does the degree of individual variation increase in depauperate assemblages?* The American Naturalist **172**, 868 (2008).

B. C. Lister, *The nature of niche expansion in West Indian Anolis lizards I: ecological consequences of reduced competition,* Evolution , 659 (1976).

J. Soininen, J. Heino, J. Lappalainen, and R. Virtanen, *Expanding the ecological niche approach: Relationships between variability in niche position and species richness,* Ecological Complexity **8**, 130 (2011).

G. Laudanno, B. Haegeman, and R. S. Etienne, *Additional analytical support for a new method to compute the likelihood of diversification models,* Bulletin of mathematical biology **82**, 22 (2020b).

F. Richter, B. Haegeman, R. S. Etienne, and E. C. Wit, *Introducing a general class of species diversification models for phylogenetic trees,* Statistica Neerlandica **n/a**, 1 (2020).

R. S. Etienne and B. Haegeman, *A Conceptual and Statistical Framework for Adaptive Radiations with a Key Role for Diversity Dependence,* **180** (2012), 10.1086/667574.

M. Foote, R. A. Cooper, J. S. Crampton, P. M. Sadler, and M. Foote, *Diversity-dependent evolutionary rates in early Palaeozoic zooplankton,* , 11 (2018).

P. Jarne, M. Loreau, N. Mouquet, P. David, and V. Calcagno, *Diversity spurs diversification in ecological communities,* Nature Communications **8**, 1 (2017).

M. J. Hamilton, R. S. Walker, and C. P. Kempes, *Diversity begets diversity in mammal species and human cultures,* Scientific reports **10**, 1 (2020).

P. Kapli, Z. Yang, and M. J. Telford, *Phylogenetic tree building in the genomic age,* Nature Reviews Genetics , 1 (2020).

M. A. Tanner and W. H. Wong, *The calculation of posterior distributions by data augmentation,* Journal of the American statistical Association **82**, 528 (1987).

K. Chan and J. Ledolter, *Monte carlo em estimation for time series models involving counts,* Journal of the American Statistical Association **90**, 242 (1995).

P. W. Glynn and D. L. Iglehart, *Importance sampling for stochastic simulations,* Management Science **35**, 1367 (1989).

D. A. Van Dyk and X.-L. Meng, *The art of data augmentation,* Journal of Computational and Graphical Statistics **10**, 1 (2001).

J. H. Friedman, *On bias, variance, 0/1—loss, and the curse-of-dimensionality,* Data mining and knowledge discovery **1**, 55 (1997).

D. G. Kendall, *On the Generalized "Birth-and-Death" Process,* The Annals of Mathematical Statistics **19**, 1 (1948).

L. M. Kieu, *Analytical modelling of point process and application to transportation,* Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent , 385 (2018).

E. Atanassov and I. T. Dimov, *What monte carlo models can do and cannot do efficiently?* Applied Mathematical Modelling **32**, 1477 (2008).

G. Celeux, D. Chauveau, and J. Diebolt, *On stochastic versions of the EM algorithm,* (1995).

T. Rydén *et al., Em versus markov chain monte carlo for estimation of hidden markov models: A computational perspective,* Bayesian Analysis **3**, 659 (2008).

J. Wang, *Em algorithms for nonlinear mixed effects models,* Computational statistics & data analysis **51**, 3244 (2007).

E. Kuhn and M. Lavielle, *Coupling a stochastic approximation version of EM with an MCMC procedure,* ESAIM: Probability and Statistics **8**, 115 (2004).

S. B. Hedges, J. Marin, M. Suleski, M. Paymer, and S. Kumar, *Tree of life reveals clock-like speciation and diversification,* Molecular biology and evolution , msv037 (2015b).

R. S. Etienne and M. E. F. Apol, *Estimating speciation and extinction rates from diversity data and the fossil record,* Evolution: International Journal of Organic Evolution **63**, 244 (2009).

G. U. Yule, *A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS,* Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character **213**, 21 (1925).

E. Paradis, *Asymmetries in phylogenetic diversification and character change can be untangled,* Evolution: International Journal of Organic Evolution **62**, 241 (2008).

J. Ng and S. D. Smith, *How traits shape trees: new approaches for detecting character state-dependent lineage diversification,* , 1 (2014).

D. L. Rabosky and E. E. Goldberg, *Model inadequacy and mistaken inferences of trait-dependent speciation,* Systematic biology **64**, 340 (2015).

S. Nee, *Birth-death models in macroevolution,* Annu. Rev. Ecol. Evol. Syst. **37**, 1 (2006).

S. Louca and M. W. Pennell, *Extant timetrees are consistent with a myriad of diversification histories,* Nature **580** (2020), 10.1038/s41586-020-2176-1.

G. Laudanno, B. Haegeman, D. L. Rabosky, and R. S. Etienne, *Detecting lineage-specific shifts in diversification: A proper likelihood approach,* Systematic Biology (2020c).

D. L. Rabosky, *Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees,* PloS one **9**, e89543 (2014).

S. Höhna, W. A. Freyman, Z. Nolen, J. P. Huelsenbeck, M. R. May, and B. R. Moore, *A bayesian approach for estimating branch-specific speciation and extinction rates,* bioRxiv , 555805 (2019).

O. Maliet, F. Hartig, and H. Morlon, *A model with many small shifts for estimating species-specific diversification rates,* Nature ecology & evolution **3**, 1086 (2019).

A. B. Phillimore and T. D. Price, *Density-dependent cladogenesis in birds,* PLoS Biol **6**, e71 (2008).

A. O. Mooers, L. J. Harmon, M. G. B. Blum, D. H. J. Wong, and S. B. Heard, *Some models of phylogenetic tree shape,* Reconstructing Evolution: New Mathematical and Computational Advances , 147 (2007).

A. Purvis, S. A. Fritz, J. Rodríguez, P. H. Harvey, and R. Grenyer, *The shape of mammalian phylogeny: Patterns, processes and scales,* Philosophical Transactions of the Royal Society B: Biological Sciences **366**, 2462 (2011).

K.-T. Shao, *Tree balance,* Systematic Zoology **39**, 266 (1990).

J. W. Fox, *Interpreting the 'selection effect' of biodiversity on ecosystem function,* Ecology letters **8**, 846 (2005).

M. Olave, L. J. Avila, J. W. Sites, and M. Morando, *How important is it to consider lineage diversification heterogeneity in macroevolutionary studies? lessons from the lizard family liolaemidae,* Journal of Biogeography **47**, 1286 (2020).

E. Bairey, E. D. Kelsic, and R. Kishony, *High-order species interactions shape ecosystem diversity,* Nature communications **7**, 1 (2016).

F. Roy, M. Barbier, G. Biroli, and G. Bunin, *Complex interactions can create persistent fluctuations in high-diversity ecosystems,* PLoS computational biology **16**, e1007827 (2020).

O. G. Pybus and P. H. Harvey, *Testing macro-evolutionary models using incomplete molecular phylogenies,* Proceedings of the Royal Society B: Biological Sciences **267**, 2267 (2000).

J. A. Fordyce, *Interpreting the γ statistic in phylogenetic diversification rate studies: a rate decrease does not necessarily indicate an early burst,* PLoS One **5**, e11781 (2010).

A. L. Pigot, A. B. Phillimore, I. P. Owens, and C. D. L. Orme, *The shape and temporal dynamics of phylogenetic trees arising from geographic speciation,* Systematic biology **59**, 660 (2010).

D. H. Colless, *Review of phylogenetics: the theory and practice of phylogenetic systematics,* Systematic Zoology **31**, 100 (1982).

T. M. Coronado, A. Mir, F. Rosselló, and L. Rotger, *On sackin's original proposal: the variance of the leaves' depths as a phylogenetic balance index,* BMC bioinformatics **21**, 1 (2020).

F. Mazel, T. J. Davies, L. Gallien, J. Renaud, M. Groussin, T. Münkemüller, and W. Thuiller, *Influence of tree shape and evolutionary time-scale on phylogenetic diversity metrics,* Ecography **39**, 913 (2016).

A. O. Mooers and S. B. Heard, *Inferring evolutionary process from phylogenetic tree shape,* The quarterly review of Biology **72**, 31 (1997).

J. B. Slowinski and C. Guyer, *Testing the stochasticity of patterns of organismal diversity: an improved null model,* The American Naturalist **134**, 907 (1989).

B. F. Oliveira, B. R. Scheffers, and G. C. Costa, *Decoupled erosion of amphibians' phylogenetic and functional diversity due to extinction,* Global Ecology and Biogeography **29**, 309 (2020).

G. R. Jones, *Tree models for macroevolution and phylogenetic analysis,* Systematic biology **60**, 735 (2011).

P. E. Smaldino, A. Lukaszewski, C. von Rueden, and M. Gurven, *Niche diversity can explain cross-cultural differences in personality structure,* Nature Human Behaviour **3**, 1276 (2019).

T. Nyman, *To speciate, or not to speciate? resource heterogeneity, the subjectivity of similarity, and the macroevolutionary consequences of niche-width shifts in plant-feeding insects,* Biological Reviews **85**, 393 (2010).

H. Robbins and S. Monro, *A stochastic approximation method,* The annals of mathematical statistics , 400 (1951).

T. Chen, E. B. Fox, and C. Guestrin, *Stochastic gradient Hamiltonian Monte Carlo,* 31st International Conference on Machine Learning, ICML 2014 **5**, 3663 (2014), arXiv:1402.4102 .

T. J. Hastie and R. J. Tibshirani, *Generalized additive models*, Vol. 43 (CRC press, 1990).

J. P. Huelsenbeck, *Performance of phylogenetic methods in simulation,* Systematic biology **44**, 17 (1995).

S. Durrleman and R. Simon, *Flexible regression models with cubic splines,* Statistics in medicine **8**, 551 (1989).

S. Wood and M. S. Wood, *Package 'mgcv',* R package version **1**, 29 (2015).

G. Nürnberger and F. Zeilfelder, *Developments in bivariate spline interpolation,* Journal of Computational and Applied Mathematics **121**, 125 (2000).

L. Meier, S. Van de Geer, P. Bühlmann, *et al., High-dimensional additive modeling,* The Annals of Statistics **37**, 3779 (2009).

J. Schaffer, *What not to multiply without necessity,* Australasian Journal of Philosophy **93**, 644 (2015).

B. C. Carstens and L. L. Knowles, *Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from melanoplus grasshoppers,* Systematic Biology **56**, 400 (2007).

C. Wiuf, *Some properties of the conditioned reconstructed process with Bernoulli sampling,* Theoretical population biology **122**, 36 (2018).

J. Whitfield, *Across the curious parallel of language and species evolution,* PLoS Biol **6**, e186 (2008).

H. Zhang, T. Ji, M. Pagel, and R. Mace, *Dated phylogeny suggests early neolithic origin of sino-tibetan languages,* Scientific reports **10**, 1 (2020).

N. Creanza, O. Kolodny, and M. W. Feldman, *Cultural evolutionary theory: How culture evolves and why it matters,* Proceedings of the National Academy of Sciences **114**, 7782 (2017).

D. Aldous, M. Krikun, and L. Popovic, *Stochastic models for phylogenetic trees on higher-order taxa,* Journal of mathematical biology **56**, 525 (2008).

T. Stadler and F. Bokma, *Estimating speciation and extinction rates for phylogenies of higher taxa,* Systematic biology **62**, 220 (2013).

J. Podani, *Different from trees, more than metaphors: branching silhouettes—corals, cacti, and the oaks,* Systematic Biology **66**, 737 (2017).

R. Molontay and M. Nagy, *Two decades of network science: as seen through the co-authorship network of network scientists,* in *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (2019) pp. 578–583.

D. H. Huson and D. Bryant, *Application of phylogenetic networks in evolutionary studies,* Molecular biology and evolution **23**, 254 (2006).

S. Chamberlain, D. P. Vázquez, L. Carvalheiro, E. Elle, and J. C. Vamosi, *Phylogenetic tree shape and the structure of mutualistic networks,* Journal of Ecology **102**, 1234 (2014).

V. Kunin, L. Goldovsky, N. Darzentas, and C. A. Ouzounis, *The net of life: reconstructing the microbial phylogenetic network,* Genome Research **15**, 954 (2005).

H.-J. Bandelt, *Combination of data in phylogenetic analysis,* in *Systematics and Evolution of the Ranunculiflorae* (Springer, 1995) pp. 355–361.

M. Farajtabar, Y. Wang, M. Gomez-Rodriguez, S. Li, H. Zha, and L. Song, *Coevolve: A joint point process model for information diffusion and network evolution,* The Journal of Machine Learning Research **18**, 1305 (2017).

K. Schliep, A. A. Potts, D. A. Morrison, and G. W. Grimm, *Intertwining phylogenetic trees and networks*, Tech. Rep. (PeerJ Preprints, 2016).

# ACKNOWLEDGEMENTS

I could write pages of acknowledges to both of my supervisors, Ernst Wit and Rampal Etienne. I was extremely lucky to end up in such an exciting project which involves two quite different research fields (macroevolutionary biology and statistical network sciences) with two of the most expert people in both fields. Professionally, I was always inspired by their way of thinking. I am grateful for the trust, space, tools, and independence they gave to me to express ideas and create this work together. Personally, I am also really grateful for their kindness and support. They are highly respected and admired by everyone around them. I will always be grateful for having them as my teachers.

I am grateful for my colleagues, always with a cooperative spirit. I had the wonderful option to go to field work in the Bornean Rainforest and feel the real experimental side of biology and data collection. The Bornean rainforest gave me a different perspective that I would never get in front of a computer. For a mathematician like me, it was a challenge to be on an excursion among biologists and I feel very grateful for all the learning I got from this group and the environment. I thank Kasper Hendriks, the expedition leader, for giving me the opportunity to travel with them, it was an experience that will be with me forever.

I also thank my friend Jorge Peña for designing the covers of this thesis and Daniela Ortega for drawing the species of figure 3.8.

I am deeply grateful for the unconditional love and strength that Cleme and Manu, my sons, gave to me every day during the journey. Always happy every day when I arrive home, embracing me with a warm hug and a lot of love.

I thank my mother for giving me the tools to survive, adapt and be happy.