# Synthesis

Richèl J.C. Bilderbeek[1]

[1]Groningen Institute for Evolutionary Life Sciences, University of Groningen, Groningen, The Netherlands

September 30, 2019

# 1

# SYNTHESIS

## 1.1. TODAY

THIS thesis shows the errors we make in phylogenetic inference when using a simpler tree prior than the one that generated the phylogenetic tree.

### 1.1.1. CHAPTER 2: babette

Chapter 2 shows the flexibility of babette, where this whole thesis shows its robustness: all the following chapters use babette to do Bayesian inference in a scripted way. Also, babette has passed the stringent code peer-review by rOpenSci, which is another indicator of its quality.

Nevertheless, there is one trait of recognition of babette missing: it has zero citations up until now, when excluding the articles in this thesis. We think the reason for it is that babette is not on CRAN yet. There are multiple reasons why this is the case: First, babette consists out of five packages. Three out of these five packages have a dependency on the others. Therefore, the packages can be submitted in sequence only. Second, babette has remained in active development since its conception. There was a prioritization of fixing a known bug and adding a vital new feature over pushing babette to CRAN and hurting users. Third, babette only supports a certain percentage of BEAST2 use cases, which are the use cases needed for this thesis.

So although babette is still underappreciated, it has only gotten better before it will get the attention it deserves. For example, after the publication of its article, a feature has been added to run babette in parallel jobs on the Groningen computer cluster, paving the way for heavy-duty usage. Without this new feature, chapters 4 and 5 would have been impossible to complete.

### 1.1.2. CHAPTER 3: pirouette

Chapter 3 shows the usage of pirouette, a tool to measure the impact/relevance of a novel tree prior. pirouette is flexible and robust enough to be used in the next two chapters. Also, when averaging out the stochasticity, pirouette works as expected, successfully quantifying the impact that a novel tree prior has.

pirouette provides a good first step to determine if a novel tree prior is relevant in a Bayesian analysis: instead of following intuition, pirouette expresses the added value of a new tree prior in numbers. The second step, however, is still missing: a clear-cut interpretation of these numbers.

* How often are new tree priors created? * Isn't it easier to write the new tree prior in BEAST2 to assess its relevance? * Pipeline is complex: * Effect of alignment's site and clock model * Effect of error function: nLTT, or gamma statistic or * Quantification still has to interpretated

* pirouette only uses an alignment and discards other data

### 1.1.3. CHAPTER 4: RAZZO

Chapter 4 shows the error in a Bayesian inference on phylogenies in which speciation co-occurs, when using standard tree priors that do not allow for co-occuring speciation. Me and my co-author quantify the errors made under a wide range of parameter settings. The more frequent to co-occurance of speciation, [it is expected] the larger the error. [it is yet unknown if that error is big compared to the background noise]

* More fine-grained parameter space, but why?

### 1.1.4. CHAPTER 5: RAKET

Chapter 5 shows the error in a Bayesian inference on phylogenies in which speciation takes time, when using standard tree priors that do not allow for speciation taking time. I quantify the errors made under a wide range of parameter settings. The longer speciation takes, [it is expected] the larger the error. [it is yet unknown if that error is big compared to the background noise] The effect of sampling a representative incipient species to create a species tree are [expected to be] small.

This thesis gives the field of phylogenetics tools and examples of how to quantify the effect of a tree prior on Bayesian inference. It will help research find the border between when speciation models are simple enough, yet not too simple. As all articles within this thesis, this thesis itself and all its source code is free (as in freedom), there is little in the way of this research contributing to the field.

## 1.2. FUTURE

* Improve babette * Apply pirouette to standard models * Measure effect of n taxa * Measure effect of DNA sequence length * Measure effect of stochasticity in alignment * Measure effect of absolute nLTT statistic: squared nLTT, log-transformed nLTT, delta R * Measure effect of more candidates * Effect of using the MRCA prior * Interpretation of error distributions * Measure impact of MBD prior differently, by adding it to BEAST2 anyways, then do an MCMC that switch models * Measure impact of PBD prior differently, by adding it to BEAST2 anyways, then do an MCMC that switch models