Synthesis

Richèl J.C. Bilderbeek[1]

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

October 6, 2019

1

SYNTHESIS

1.1. PAST

THIS thesis started with two very basic biological questions. The first question is to ask which species are closest related to one another. This can be answered satifactorily nowadys using DNA sequences: the taxa of which the DNA sequences is most similar, are closest related.

The second very basic biological question, is to ask *when* these speciation events took place. This thesis shows `babette` (chapter 2), a tool that allows to create a dated phylogeny from DNA sequences.

Constructing *a* dated phylogeny from a DNA alignment using *a* tool is easy. It is harder to get a reasonably accurate dated phylogeny, because one needs to make a defendable pick from the many combinations of models. To make a defendable pick, however, we can sometimes use a rational procedure. With `mcbette` (a component of `pirouette`, chapter 3), one can do a model comparison, which concludes which inference model is best picked on an alignment.

Even would we know the model that nature indeed follows, we still make an error in constructing our dated phylogeny: due to chance, the resulting DNA alignment may have too less or too much mutations in certain places.

This brings us to the bigger question underlying this thesis: How well can we construct a phylogeny from an alignment? With theoretical studies, in which we have complete knowledge, we can measure the error in the phylogenies constructed.

One thing we do not know, is what is the true phylogetic model in nature. A phylogenetic model is a combination of models that each highlights a part of the process. In this thesis I focus on the speciation model, of which, we do not know what is the true speciation model in nature.

Due to this, there are many speciation models, each incorporating a facet of the speciation process. When we have to pick a defensible speciation model, we are usually pragmatic and pick a model that is incorporated within our phygenetic tools. There are cases, however, in which we could worry that these standard models may be overly simplistic, because they overlook a facet of speciation that may be very convincing in some cases.

In this thesis, me and Giovanni Laudanno, created a tool to measure the error we make in inferring a phylogeny, due to picking the wrong speciation model. With this novel tool, called `pirouette` (chapter 3), the field of phylogenetics now has a method, with which we demonstrate if and when the standard speciation models are justifiably good.

There are two speciation models that are (yet) non-standard. The first speciation model, unlike the standard speciation models, allows speciation to co-occur. When there is a scenario in nature, in which a process triggers speciation in multiple species, this multiple-birth death (MBD) model would be the better fit. In chapter 4, Giovanni Laudanno and me use `pirouette` to measure the error we make, when we use a standard speciation model on a process we know (read: simulated) to follow the MBD process. From that we found that ...

The second non-standard speciation model, unlike the standard speciation models, assumes that speciation takes time, which is encorporated in the protracted-birth death (PBD) model In chapter 5, I use `pirouette` to measure the error we make, when we use a

standard speciation model on a process we know to follow the PBD process. From that we found that ...

1.2. OPEN QUESTIONS AND FUTURE WORK

Now we can measure the errors we make in our phylogentic inference, when using a standard phylogentic model on a true process following a non-standard model. `pirouette` can serve as a litmus test to measure the relevance of a novel tree prior.

Of course, there are already ways to estimate the relevance of a novel speciation model. A straighforward one, is to add the speciation model to the set of standard models, after which its inference is compared with the standard models. A drawback of this approach, is that it is harder to develop: not only need the novel speciation model be able to simulate its phylogenies, also a likelihood equation is needed to allow it to be used by the phylogenetic tools. Developing such a likelihood equation may be harder than using `pirouette`.

The step forward of `pirouette` is that it allows to quantify the error we make in our phylogentic inference, by expressing it as (usually) two distributions: one with the baseline errors, one with the added error caused by using an incorrect phylogenetic model. What is missing in `pirouette`, is a clear-cut interpretation of these error distributions, that is, a clear yes/no answer to the question: 'Is using the standard phylogenetic models good enough?'. Only when the two distributions overlap, can we confidently claim a yes.

There have been defensible, yet arbitrary, choices made in the pipeline of `pirouette`. For example, a twin alignment has as much mutations acculated from the root sequence, as the true alignment. This design choice should assure the Bayesian inference in the next step to work on an equal amount of genetic information. Unknown is if this indeed improves the judgements made by `pirouette`. These choices should one day be parameters of using `pirouette` as well.

We apply the `pirouette` methodology in chapter 4 and 5. We show that the error in a Bayesian inference on phylogenies increases when the effects of co-occuring or protracted speciation are stronger. This qualitative conclusion is obvious, but the extent of these errors has never been quantified before. An obvious follow-up question is, what are examples in nature where we expect a high abundance of co-occurring or protracted speciation, and thus, what is the error empiricist make?

Disregarding the results of chapters 3 and 4, there are at least two reasons to add the MBD and PBD tree models to the standard phylogentic models anyway. The first reason is methodological: perhaps the `pirouette` setup gave rise to an inrepresentative conclusion. Using the actual tree prior in inference and comparing its performance to the standard ones, is still the most straightforward way to determine, ironically, if the novel tree prior adds value as a standard tree prior. The second reason to add a novel tree prior the standard models, is because of the model parameters that are estimated jointly with the phylogenies. The clearest example is the PBD model, which allows one to estimate biologically relevant parameters such as the duration of speciation.

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.