Introduction

Richèl J.C. Bilderbeek[1]

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

October 21, 2019

1

INTRODUCTION

[ALP: I feel the two aspects of the introduction that need the most work are 1) better explaining why it is so important that we can accurately reconstruct phylogenetic relationships 2) better explaining why current models may not be adequate and what new models need to be developed. Just saying that current models assume 'x' but this may not be appropriate if the real process is 'y' but we don't know 'y', is not enough. Instead, you really need to delve into the speciation literature, discussing the theory and empirical evidence for how the speciation process occurs, illustrating your discussion with real world examples. Have you read 'Speciation' by Coyne and Orr and also 'Speciation in Birds' by Trevor Price? These two books are excellent.]

1.1. SPECIATION

S PECIATION is the process that creates new species, connecting all of life to one shared common ancestor. It is a process that has resulted in the millions of species on Earth nowadays, as well as in the many species that have gone extinct. Some speciation events that gave rise to extant species, happened earlier than others, from hundreds of millions of years ago (so-called 'long-enduring species', or, informally, 'living fossil') to more recent ones. See figure 1.1 shows an example of each.

Figure 1.1 | An long-enduring species (left) and a young species (right). The species at the left is a preserved specimen of *Latimeria chalumnae*, estimated to exist for hundreds of millions of year. The species at the right is the *Homo sapiens*, existing for around a third of a million years.

1.2. RELATEDNESS

All species on Earth are related, as all organisms are derived from a same universal common ancestor. It is estimated that the last universal common ancestor (LUCA) lived around 3.5 (Doolittle 2000), to 4.5 billion years ago.

1.2.1. INTRODUCTION

A first very basic question within the field of biology, is to ask which species are closest related to one another. Take, for example, the 8 crocodilian species depicted in figure 1.2.

1.2.2. ANSWER WITH MORPHOLOGY

To

1.2.3. ANSWER WITH DNA

1.2.4. DIFFERENCE BETWEEN MORPHOLOGY VERSUS DNA

You may come up with one or more pairs of hypothesized closest-related species, but you will probably get stuck on which ancestors of these pairs are most closely related. Note that doing the other way around, to use morphology to classify species, is tricky: how to decide when the morphology between two specimens is different enough to classify these as two species?
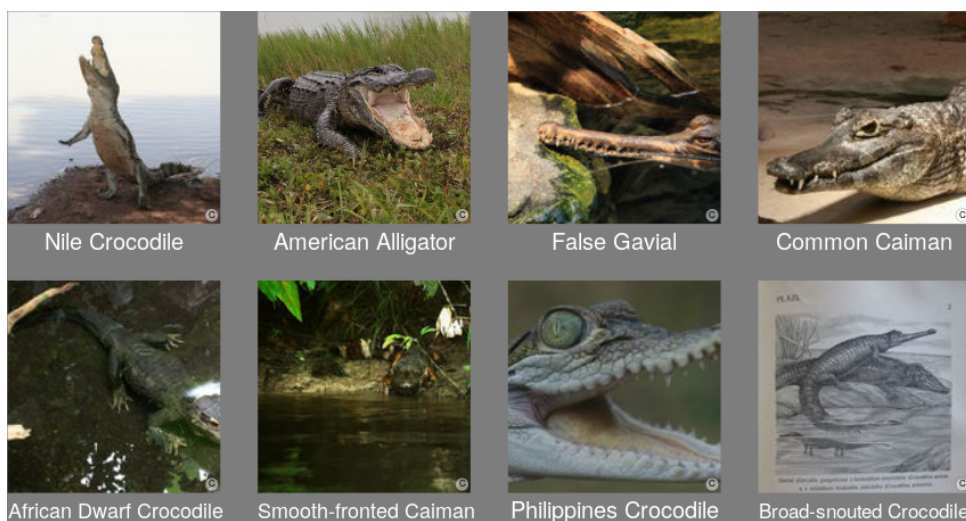

Figure 1.2 | Eight members of the order of crocodilians

1.2.5. RELATEDNESS: BIOLOGICAL RELEVANCE

Knowing which species are closest related to one another goes deeper than just being able to do taxanomic bookkeeping. Each speciation event, gives us a hint in understanding speciation. From two sister species, we can try to deduce the cause if their speciation. It could be that the ancestor species got geographically isolated and that this would be a major cause of speciation (e.g. Mayr, 1942). But the reason for that speciation event may have also been ecological (e.g. Lack, 1947), for example, by the ancestral species evolves from a generalist species into to specialist species. Also, are some species (and their descendants) likelier to speciate, and if yes, what causes that?

1.2.6. RELATEDNESS: TOOLS

[RJCB: Describe early computational tools to create cladograms here]

1.3. TIME OF SPECIATION

1.3.1. TIME OF SPECIATION: INTRODUCTION

The second very basic biological question, is to ask *when* these speciation events took place.

1.3.2. TIME OF SPECIATION: BIOLOGICAL RELEVANCE

Also this question goes deeper than just adding a timescale to phylogenies. Knowing when a speciation event took place, opens up many clues in understanding speciation, as, for example, that moment in time may be linked to a large geographical change. Additionally, we can (try to) deduce the species community through time. Not only does speciation influence species communities (by adding new members), a species community might influence the process of speciation in return. An open question is whether species communities get saturated to a maximum number of species, or that new community members give rise to new niches for more new members.

1.3.3. TIME OF SPECIATION: ANSWER WITH MORPHOLOGY FAILS

This question cannot be answered based on morphologies of the present-day species alone, because morphology is a complex trait, and the pace at which morphology changes in time is unknown or unpredictable.

1.3.4. TIME OF SPECIATION: ANSWER WITH FOSSILS IS OK

This second question can be answered by using a classical approach, by using the morphology of fossils. This approach can only be used if the species *can* fossilize, and those fossils are found in multiple points in time. Even if this is the case, there are caveats. Using morphology on extinct species is even trickier, as species change their appearance in time. Also an imaginary time machine would not help us out: we could try to determine the number of species in each timepoint, but that would only work if we could confidently define what a species is. We cannot, because speciation is usually a gradual process.

1.3.5. TIME OF SPECIATION: ANSWER WITH DNA IS AWESOME

This second question can also be answered using a modern approach, by using the DNA sequences of extant species, as shown, for example, in figure 1.3. Because DNA is inherited from parent to offspring and changes through times, it carries each species' evolutionary histories within it. The point in time when a species speciates is marked by the two daughter species having seperate mutations from that moment on. Due to this, we can easily find closest related species by measuring the similarity in DNA sequences. If we know how frequent mutations occur, we can already do a rough estimation of when the speciation event took place. In reality, DNA sequences of different species varies in length, due to insertions and deletions in genetic sequences, but in the simulation studies in this thesis, we will ignore this.
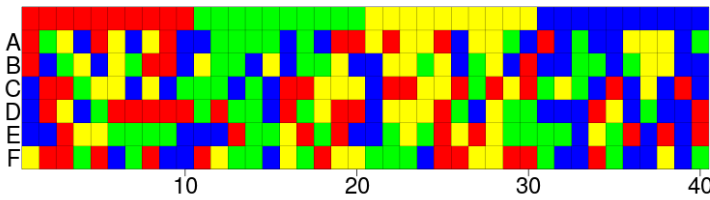
Figure 1.3 | A 40-nucleotide DNA alignment of six hypothetical species. The species are named A to and including F. The four colors denote the four different nucleotides, in which the red color resembles adenine, yellow depicts cytosine, green is for guanine, and blue resembles thymine. The top row shows the (artificial) root sequence, which is usually unknown.

1.4. PHYLOGENETICS

The field of phylogenetics is devoted to use heritable information of (usually) extant species to infer a dated phylogeny. Phylogenetics is a field that has made enourmous leaps thanks to the increase of computational power. Phylogenetics allows us to follow the evolution and distribution of traits and behavior in time, and -due to this- understand the evolution and distribution of species diversity. All of this depends on an accurate reconstruction the species' dated phylogenies.
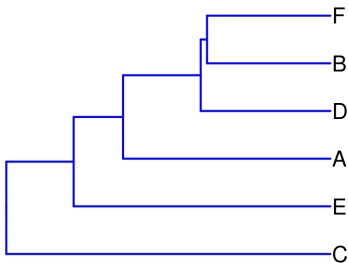

Figure 1.4 | Phylogeny created from the alignment in figure 1.3 using a quick-and-dirty methodology.

1.4.1. NON-BAYESIAN PHYLOGENETICS: SIMPLE EXAMPLE

There is a rich toolset to infer a phylogeny from heritable information (which, in this thesis, always consists of a DNA alignment). Figure 1.4 shows a phylogeny inferred from the alignment in figure 1.3 using a quick-and-dirty methodology (of which the name is irrelevant). The phylogeny shows the six hypothetical species and their evolutionary relationships. Going from left to right, we travel through time from the past to the present. The leftmost vertical line indicates the first speciation event, which gave rise to the first two ancestral species. This first split in the tree is called the crown, the moment in time this occurred is called the crown age.

1.4.2. NON-BAYESIAN PHYLOGENETICS: PIPELINE

The problem with phylogenies is, that it is impossible to go out in the field and measure one, as they depict which species lived when *in the past*. Instead, we *construct* phylogenies. For example, the phylogeny in figure 1.4, how well does match the true phylogeny? **That question, is the main question of this thesis: how well can we construct a phylogeny from an alignment?** What is the error we make when we construct a phylogeny?

1.4.3. NON-BAYESIAN PHYLOGENETICS: PIPELINE SIMPLE EXAMPLE

Answering this research question is, at first glance, easy:

1) simulate a true phylogeny.

2) simulate an alignment that follows that phylogeny.

3) construct a phylogeny from that alignment.

4) measure the difference between the true and constructed phylogeny.

This workflow is depicted in figure 1.5. All steps, however, are more complex than just this.
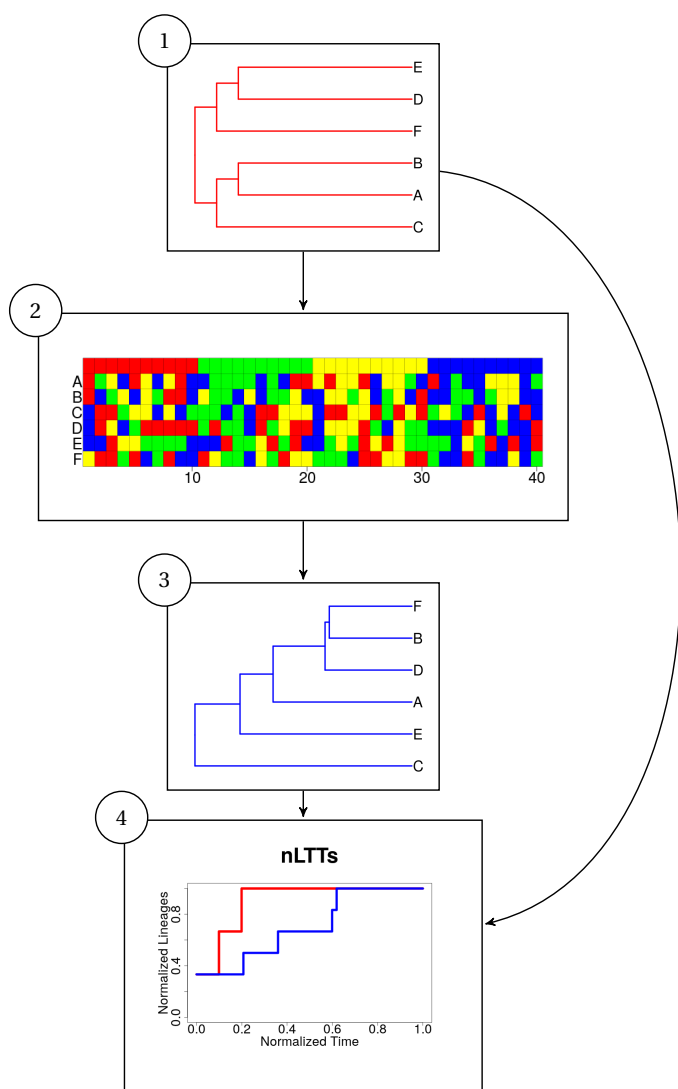
Figure 1.5 | (Simplified) method to answer the research question of this thesis: 1. simulate a true phylogeny. 2. simulate an alignment that follows that phylogeny. 3. construct a phylogeny from that alignment. 4. compare the true and constructed phylogeny. Note the big difference between the true and constructed phylogeny

1.5. BAYESIAN PHYLOGENETICS

1.5.1. INTRODUCTION

Constructing a phylogeny from an alignment is the step that gets most attention in this thesis, as it is also the most complex one. Unlike the methods described earlier, we do not construct one single phylogeny, but we construct a distribution of multiple phylogenies.

Within this distribution of multiple phylogenies, the phylogenies that are more likely, will be present more often. This method is called Bayesian phylogenetics, in which we use a Bayesian approach to create phylogenies based on genetics.

1.5.2. BEAST2

1.5.3. BABETTE

To be able to do the phylogentic inference needed for the rest of this thesis, I developed an R package to do so, which is discussed in chapter 2. Because a Bayesian inference can be set up in many ways, the greatest asset of that package is that it gives a consistent grammar to express each setup. Additionally, the R package allows to run Bayesian inference from the command-line, which is essential for the theoretical studies in this thesis.

If we can measure the error we make in our inference, we can try and improve the inference. One way to improve it, is to use a better inference model. Ideally, we would use the same inference model that gave rise to the true phylogeny and alignment, but, alas, we (usually) do not know that model. We do not know that model, because we do not know the model that nature used: the processes that cause speciation are possibly many, and the mechanisms of each are unknown and/or debatable.

[ALP: ALP: but you need to go into this in detail to build the case for why the existing models are unlikely to be sufficient or at least why it is important to explore potentially more realistic models]

Due to this, we'll have to resort to the many phylogentic models to explain the DNA (RNA, protein, morphological and fossil) data best.

There are plenty of phylogentic models, ranging from simplistic to very complex. The most popular models make it into our phylogenetic programs (which, in turn, may make these models even more popular), which I will define as 'standard models'. When empiricist build a phylogenetic tree from their painstakingly acquired DNA alignment, they pick their favorite standard model or use an algorithm to select one. The empiricist assumes that the standard models are good enough for his/her cause.

The standard models, however, make some assumptions that will not hold in all biological cases.

[ALP: ALP: expand on this to explain what the standard model is and what assumptions they make]

This will increase the error we make in our inference. But will that error be profound enough to reject using a standard model?
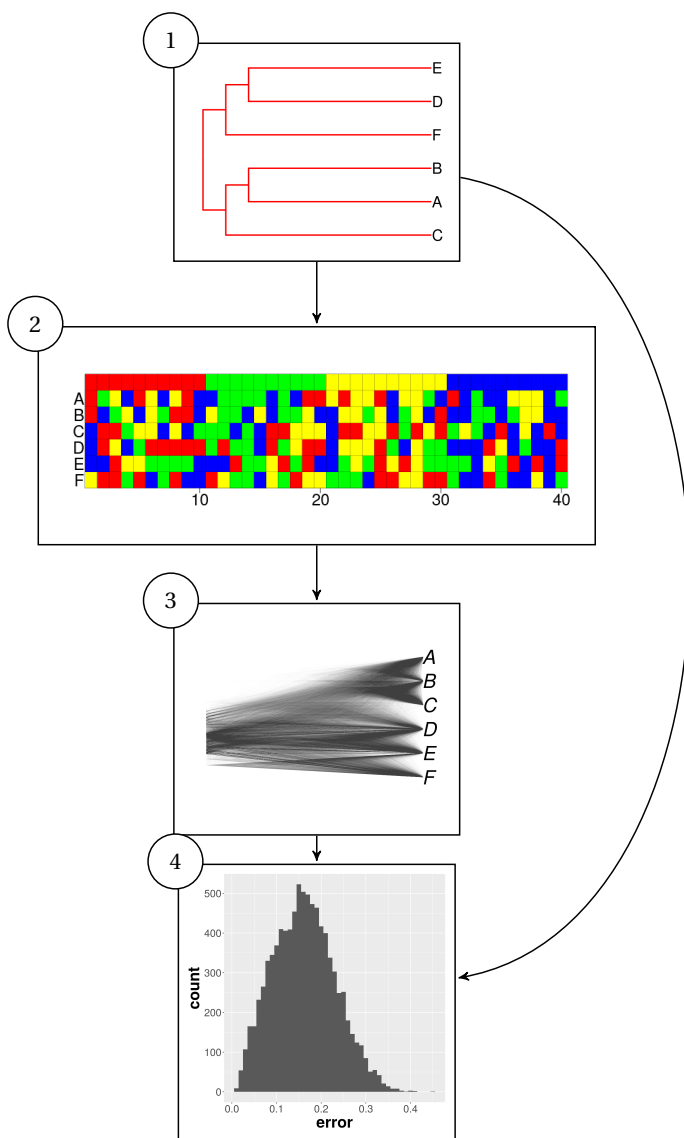
Figure 1.6 | Method to answer the research question of this thesis: 1. simulate a true phylogeny. 2. simulate an alignment that follows that phylogeny. 3. infer a distribution of phylogenies from that alignment. 4. compare the true phylogenies with the inferred phylogenies.

1.5.4. STANDARD ASSUMPTIONS

With Bayesian inference, we need an alignment and our model assumptions to infer phylogenies. Different model assumptions will result in different trees. There are many assumptions to choose from, regarding the mutation of nucleotides in the DNA alignment, the DNA mutation rate of species and the speciation model. The model assumptions

specify what we assume to be true regarding how the alignment came to be.

For example, we can assume that:

1) the true phylogeny had a constant speciation and extinction rate

2) the mutation rate is constant and equal for all species

3) all mutations between nucleotides are equally likely

This set of model assumptions is simple and -party due to that- commonly used. Figure 1.7 shows an example, that, for an alignment obtained from primates, it matters little if extinction is assumed to be absent or constant.

[ALP: This is the point at the heart of these thesis, right? In order to infer phylogenetic topology and divergence times we need to assume a certain model for speciation and extinction. You need to build the case for why the current model used may not be appropriate and why more complex models may be required. You really need to delve into the 1) literature on speciation, to discus the biological/geographical/environmental mechanisms underlying speciation and our latest understanding of how speciation works 2) literature on modelling diversification, highlighting how there have been major advances in modelling different modes and tempo of speciation when estimating diversification but not (surprisingly) when infering trees. You need to highlight the logical inconsistency of this and why it could be severely problematic - if we assume certain speciation modes when infering trees does this bias what speciation mode we would infer from that tree??]

[RJCB: Although I think this feedback is excellent, I will prioritize handing in a thesis chapter approximately on time, therefore ignoring most of this feedback for now]
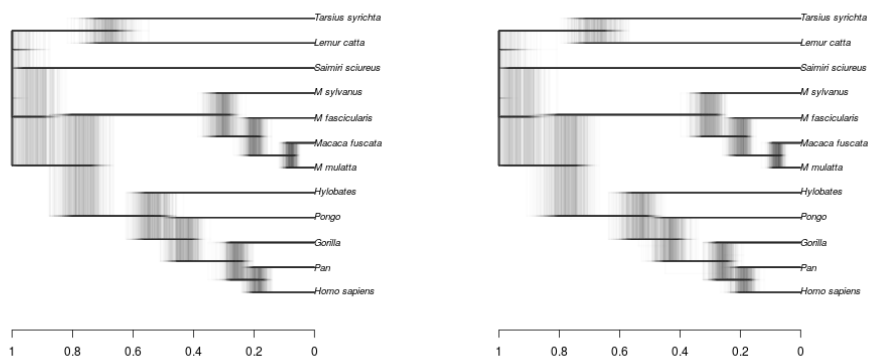

Figure 1.7 | Phylograms created from the same alignment, using three different speciation models: left: Yule, right: Birth-Death. Both timescales are normalized to one. Note that for both speciation models, the genus *Pan* and the species *Homo sapiens* separated at approximately the same time.

The result of a Bayesian inference, is -to be precise- a posterior disribution of jointly-inferred phylogenies and model parameter estimates, simply called 'posterior' in this thesis. The way such a posterior is generated assures that more likely phylogenies are present more often. This distribution of phylogenies shows the (un)certainty of the inference. For example, the posterior phylogenies in figure 1.6, panel 3, show a high degree of uncertainty, as the inferred phylogenies vary widely in shape. The posterior correctly suggests two clades (ABC and DEF), but does not confidently show the two most related taxa (AB and DE). We can already make the rough claim that, would the phylogeny in panel 1 in figure 1.6 depict a true phylogeny, we make a big error in its inference.

1.5.5. ASSUMPTIONS MISMATCH

There are, however, plenty of biological cases where the assumption mismatch:

MBD
MBD

biology   One assumption of all standard models is that speciation event happen independently, that is, there are never two speciation events at the same time. There are biological scenario's in which may say this is false: when a habitat is split up, due to a geological barrier, this will result in two species communities. The change from one to two communities is likely to affect both communities and trigger a speciation event in both communities.

[ALP: ALP: Great! this is what i was looking for. You need much more of this kind of stuff. There are other scenarios you might want to discuss e.g. barriers that result in multiple species being isolated simultaneously like when habitat layers move up and down mountains during ice ages species get isolated on the tops or in the valleys.]

razzo   The inference error of ignoring co-occurring speciation is quantified by me and Giovanni Laudanno in chapter 4.

PBD

biology   Another assumption of all standard models is that speciation event happen instantaneously, that is, when there is a speciation event, the two species are immediately recognized as such. We know that speciation takes time.

[ALP: you need to provide a lot more information and discussion here - why does speciation take time? Does it always take time or can it be instantaneous? Are there examples of clades that you can use to illustrate your arguments? You need to delve into the theory and empirical evidence of speciation here to make a compelling case for why a protracted model that at least allows speciation to occur gradually is needed. Again, highlight how this has been implemented to infer speciation dynamics from tree but oddly the models used to infer the tree in the first place ignore this.]

raket   The inference error of ignoring this fact is quantified by me in chapter 5.

1.5.6. PIROUETTE

To be able to determine the impact of using a standard phylogentic model, when we know the biological process in more complex than it assumes, me and Giovanni Laudanno developed an R package to quantify the error we make due to this mismatch, which is described in chapter 3.

1.5.7. CONCLUSION

In chapter 6, I show which conclusions can be drawn from these chapters

1.6. REFERENCES

REFERENCES

Doolittle, W.F. (2000) Uprooting the tree of life. *Scientific American*, **282**, 90–95.

1.7. AFTER REFERENCES

[RJCB: note the comments below were made about the Synthesis-wrongly-assumed-to-be-Introduction, so these points should be in the Introduction, but these comments are not about this doc]

[ALP: First, is to take a step back and consider your work in the broader context. Why do we need to be able to accurately infer phylogenetic relationships? Why is this important? What questions rely on this? There needs to be a logical progression that sets the scene and highlights to the reader why your work is relevant e.g. we need to understand biodiversity, a key tool to do this is phylogenies, there have been major advances in methods to infer evolutionary processes but all this depends on an accurate phylogeny.] [ALP: Second, try and be more biological. The document is quite technical and methods focussed but i think you need to ensure that the biological process is forefront. It is never made clear why these new speciation models that you have implemented are more realistic than the standard models.] [ALP: Third, once you have set the scene you then need to identify the problem and propose the solution. e.g. current tree inference are based on models of the speciation process. But these are very simplistic compared to how speciation actually happens. We suspect that this may lead to biased inference and the aim of this project is therefore to.] [ALP: Fourth, once you have set the scene and presented the problem you can then discuss in more detail what each of your chapters addresses. Overall, I found the layout and titles of 'past' and 'open questions and future work' confusing e.g the second of these sections has suggestions for further work as well as description of what you did. I suggest this is re-organised.] [ALP: Finally, you need to work on having a strong conclusion that brings together all your research findings and presents your vision for the field.]

1.8. PHOTO ATTRIBUTION

Figure 1.1, the left image, Preserved specimen of chalumnae by Alberto Fernandez Fernandez is licensed under CC BY-SA 3.0. the image at the right, Akha couple in northern Thailand by Weltenbummler84 is licensed under CC BY-SA 2.0 DE.

For figure 1.2, the selection of these eight images was done by OneZoom. These images, from top-left to bottom-row, row-first: Nile Crocodile by Marco Schmidt is licensed under CC-BY-SA 3.0. American Alligator by NASA Kennedy. False Gavial by Yinan Chen is marked as being in the public domain. Common Caiman by Michael Wolf is licensed under CC-BY 2.5. African Dwarf Crocodile by Staycoolandbegood is marked as being in the public domain. Smooth-fronted Caiman by Whaldener Endo is licensed under CC-BY-SA 4.0 Phillipines Crocodile by Vanderploeg is marked as being in the public domain. Board-snouted Crocodile by Hadonos is Marked as being in the public domain.