Introduction

Richèl J.C. Bilderbeek[1]

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

September 24, 2019

1

INTRODUCTION

SPECIATION is the process that creates new species, connecting all of life to one shared common ancestor.

Figure 1.1 | Eight members of the order of crocodilians, as selected by OneZoom 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.

This process can be investigated on multiple levels, for example at the individuals' or at the species level. In this thesis, I focus on the latter.

Speciation at the species level simplifies a species to a horizontal line (also called a 'branch') in a phylogeny, answering basic questions like 'Which species lived when?', 'When did a speciation event take place?' (depicted by the vertical lines) and 'Who is the ancestor of which species?'. Figure 1.2 shows an example phylogeny:

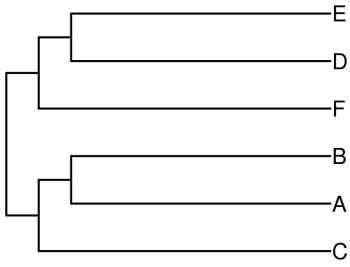


Figure 1.2 | A phylogeny with six species

Figure 1.2 shows a phylogeny (also called 'phylogenetic tree', or simply 'tree') with 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.

Each of these ancestral species gives rise to its own evolutionary history, resulting in a tree with two clades: the ABC and the DEF clades. Sure, we could equally well have started the phylogeny with one ancestral species at the utmost left, going further back in time from the crown age to, what is called, the stem age, but in the context of this thesis, we do not.

It is impossible to go out in the field and measure a phylogy, as they depict which species lived when *in the past*. Even an imaginary time machine would not help us out: we could try to count 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.

Without resorting to imaginary machines, we -at the present- can work with extant species, as they carry their own evolutionary history with them, in the form of DNA. From those species, we can obtain (part of) their DNA sequences. DNA sequences of different species may vary in length, due to insertions and deletions in genetic sequences, Usually, we use a procedure (a software tool, for example) to align the sequences, as shown in figure 1.3:

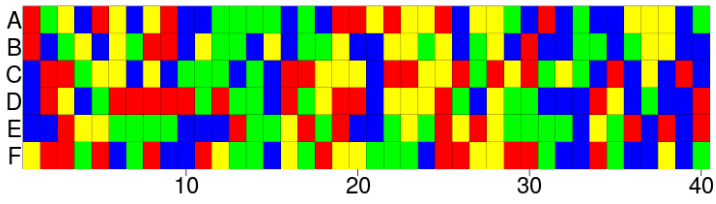


Figure 1.3 | A 40-nucleotide DNA alignment of the six species

Figure 1.3 shows an alignment of our six hypothetical species that we actually could have found in nature. From this alignment, we can *infer* a phylogeny, which basically means 'best guess following a rational procedure'. There are multiple ways to infer a phylogeny, for example, using maximum likelihood or Bayesian inference. In this thesis, I focus on the latter.

With Bayesian inference, we use an alignment and our model assumptions to infer a posterior (more precise: 'a joint posterior distribution of phylogenies and model parameters'). We do so, by first creating a random phylogeny. Using a likelihood equation, we can calculate how likeli it For that, we use a Markov Chain Monte Carlo algorithm, which is n

A posterior contains multiple inferred phylogenies, in which the more likely ones are present more often. This distribution of phylogenies shows the (un)certainty of the inference. Figure 1.4 shows the posterior phylogenies we obtain from our alignment:

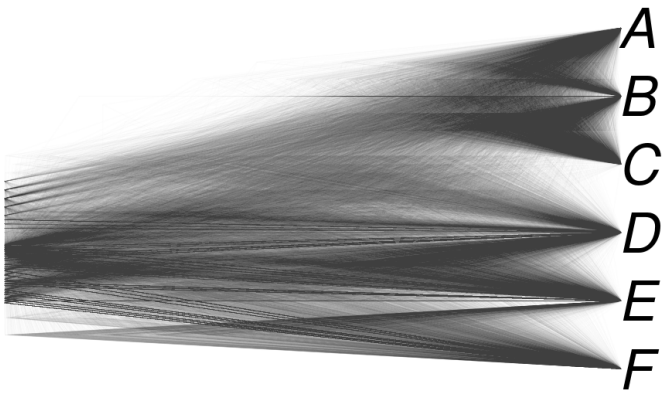


Figure 1.4 | The posterior phylogenies of the six species, from a DNA alignment of 40 nucleotides

Figure 1.4 shows a high degree of uncertainty, as the inferred phylogenies vary widely in shape. The inference only weakly distinguishes between the ABC and DEF clades.

The inference described so far is unsatisfactory, as we can only draw weak conclusions. We can improve the inference by using a longer DNA sequence or by picking a better inference model. In a simulation study, we can easily increase the number of nucleotides, figure 1.5 shows the posterior phylogenies we obtain from our longer alignment:

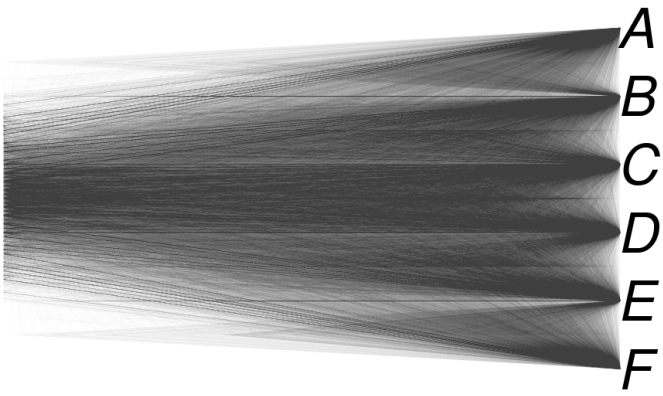


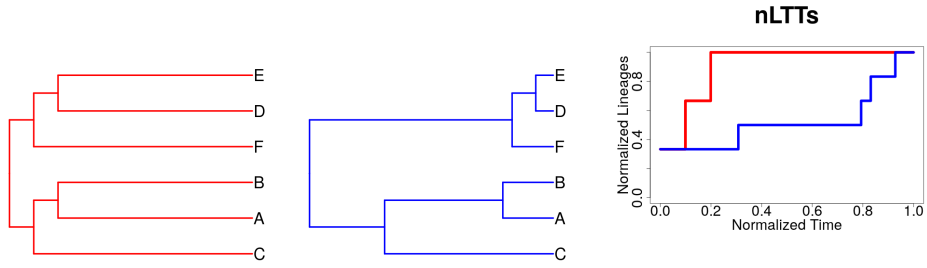
Figure 1.5 | The posterior phylogenies of the six species, from a DNA alignment of 400 nucleotides

Figure 1.5 shows that in this example, with more information, we can only show our uncertainty more clearly.

Another way to improve our inference is using a better inference model. An inference model embodies our assumptions on how we think evolution works, and consists of (1) how nucleotides mutate to others (also called 'the site model'), (2) how often mutations occur (also called 'the clock model'), and (3) how speciation works (also called 'the tree prior' or 'the speciation model'). Theory predicts that usually the inference model becomes less important if there is more information in an alignment. The example shows here, however, shows one of the exceptions.

Ideally, we pick an inference model identical to the actual way things work (or: 'the true model'), be it in nature or *in silico*. In practice, the true model that nature uses is unknown. Due to this, scientists came up with many models to explain the DNA (RNA, protein, morpholical and fossil) data best.

In a theoretical study, when can simply pick how nature works; that is, how it is simulated. Theoretical studies are useful, as these explore how well we will ever be able to explain nature. To do so, a 'true' speciation model is picked to generate 'true' phylogenies. From these phylogenies, a 'true' site model and clock model are used to simulate a 'true' alignment. From that 'true' alignment, which is the data we can gather from nature, we can then see how close our inferred phylogenies are to the 'true' phylogeny.



(a) True tree **(b)** Twin tree **(c)** Combined nLTT

Figure 1.6 | nLTTs

There are many ways to quantify how similar two phylogenies are. The normalized lineages-through-time (nLTT) statistic (**?**) simplifies a phylogeny to a number of lineages (the number of branches) in time. Both number of lineages and time are normalized to have a maximum of one, which allows us to compare two trees of any number of tips of any crown age. The difference between two phylogenies is simply the surface between the two phylogenies' nLTT plots. If two phylogenies are identical in normalized shape, the nLTT difference between them in zero, else the value will be higher, with a maximum of one. Because the value of the nLTT increases with increasing difference between the trees, the nLTT statistic is a measure of difference, or error.

In this thesis, I quantify the errors we make in our phylogenetic inference:

- In chapter 2, I show an R package I developed to do Bayesian inference from the command-line
- In chapter 3, me and Giovanni Laudanno describe an R package we developed to quantify the error we make in phylogenetic inference
- In chapter 4, Giovanni Laudanno and I quantify the error we make in phylogenetic trees when speciation can co-occur
- In chapter 5, I quantify the error we make in phylogenetic trees when speciation takes time
- In chapter 6, I show which conclusions can be drawn from these chapters

2

PHOTO ATTRIBUTION