Introduction

Richèl J.C. Bilderbeek[1]

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

September 27, 2019

1

Introduction

SPECIATION 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 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.

A first very basic question within the field of biology, is to ask which species are closest related to one another. If you think that’s an easy question, try to apply it, for example, on the 8 crocodilian pictures in figure 1.2. 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?

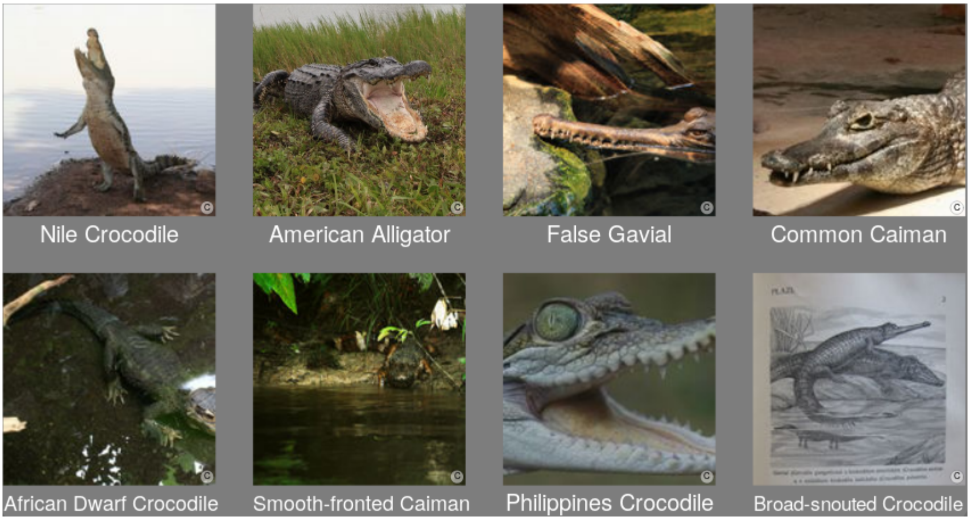


Figure 1.2 | Eight members of the order of crocodilians

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

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.

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.

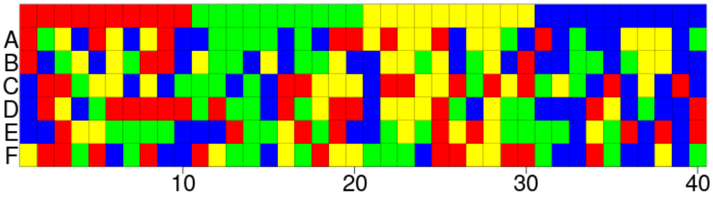


Figure 1.3 | A 40-nucleotide DNA alignment of six hypothetical species. The species are named A to and including F. The four colors denote the four different nucleotides, in which the red color resembles adenine, yellow depicts cytosine, green is for guanine, and blue resembles thymine. The top row shows the (artificial) root sequence, which is usually unknown.

Using the DNA sequences of extant species to answer our biological questions, however, is a complex topic. For starters, there are multiple ways to do so, like bootstrapping, jack-knifing, parsimony, maximum likelihood and more. A conveniently simple approach is to use UPGMA, which answer our biological questions, in the form of a phylogeny.

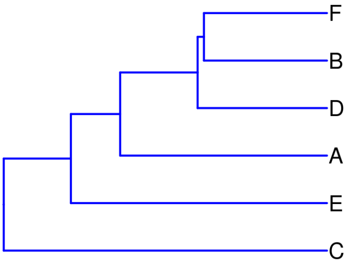


Figure 1.4 | Phylogeny created from the alignment in figure 1.3 using UPGMA. This method is irrelevant in the context of this thesis.

Figure 1.4 shows a phylogeny (also called ’phylogenetic tree’, or simply ’tree’), created from the alignment in figure 1.3 using one such methodology. It 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. 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.

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 ’tree without birds’?

Answering this research question, at first glance, is 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.

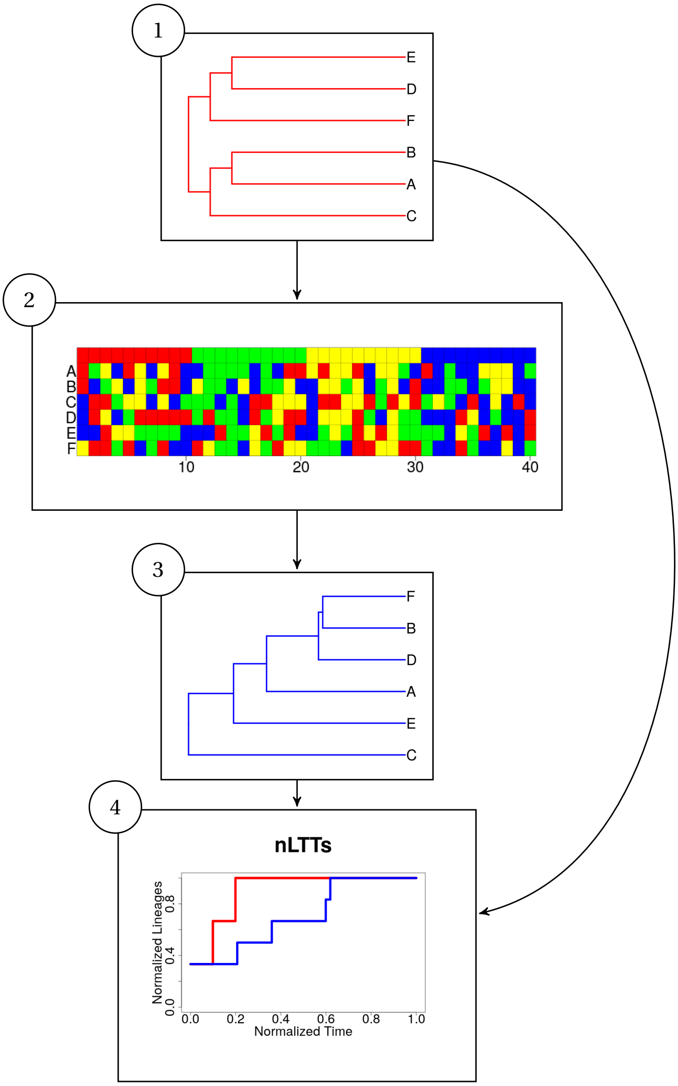


Figure 1.5 | (Simplified) method to answer the research question of this thesis: 1. simulate a true phylogeny. 2. simulate an alignment that follows that phylogeny. 3. construct a phylogeny from that alignment. 4. compare the true and constructed phylogeny.

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 1.6 shows the posterior phylogenies we obtain from our alignment:

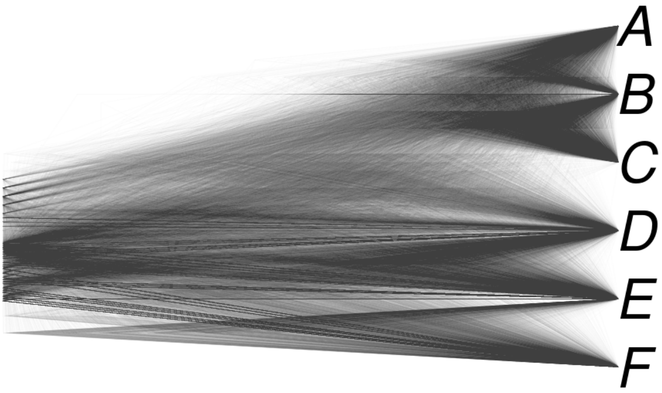


Figure 1.6 | The posterior phylogenies of the six species, from a DNA alignment of 40 nucleotides

Figure 1.6 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 1.7 shows the posterior phylogenies we obtain from our longer alignment:

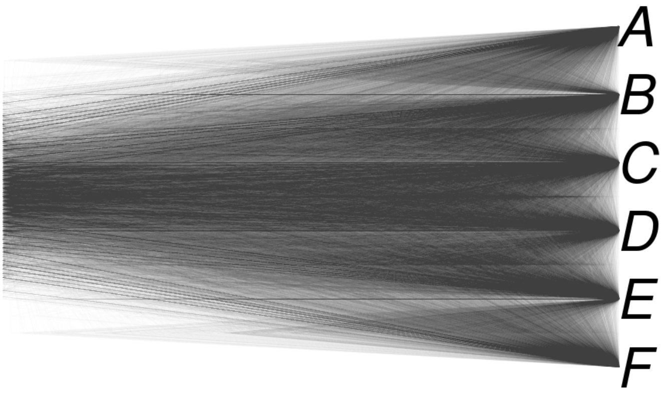


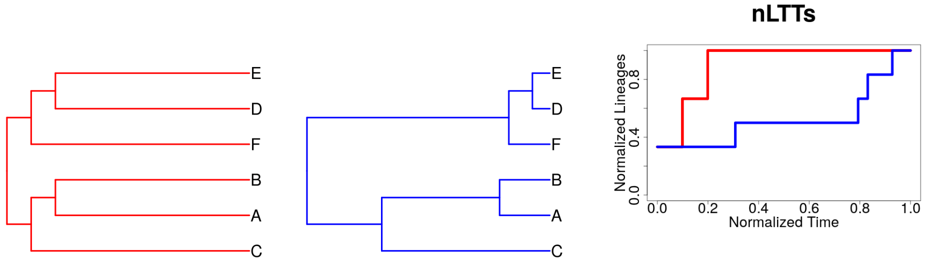
Figure 1.7 | The posterior phylogenies of the six species, from a DNA alignment of 400 nucleotides

Figure 1.7 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 1.8 | 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

Figure **??** shows an example phylogeny:

1.0.1. 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.