Introduction

Richèl J.C. Bilderbeek[1]

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

February 9, 2020

1

INTRODUCTION

Once upon a time, there was the first living organism, the First Universal Common Ancestor (FUCA). We do not know when it lived.

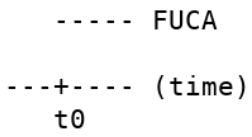

Figure 1.1 | Evolutionary history of the First Universal Common Ancestor (FUCA). Time goes from past (left) towards the present (right). **[RJCB: TODO: Use proper phylogeny with proper timescale]**

One unknown day, FUCA speciated, resulting in two species. This event doubled the biodiversity on Earth. The two species, which we will call species A and B are sister species. We do not know what caused the speciation.

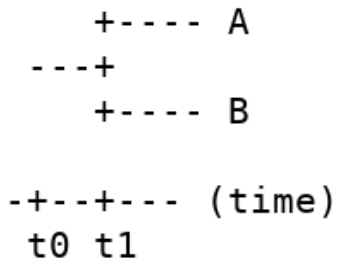

Figure 1.2 | Evolutionary history of the First Universal Common Ancestor (FUCA). and its two decendants. Time goes from past (left) towards the present (right). **[RJCB: TODO: Use proper phylogeny with proper timescale]**

Both species A had their unknown histories: they speciated themselves, and they and/or their descendants went extinct. Extinction is a common event. Let's assume A and/or its clade went exctint and that species B created a sister species C. Species B and C will give rise to all contemporary biodiversity. This ancestor of species B and C is called the Last Universal Common Ancestor and lived around [then].

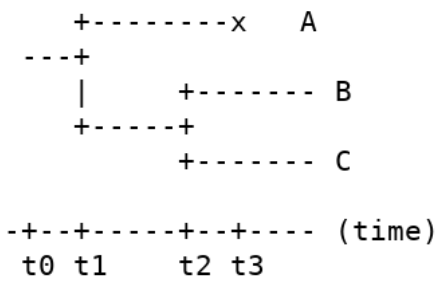

Figure 1.3 | Evolutionary history of the First Universal Common Ancestor (FUCA). and its three decendants of which one went extinct. Time goes from past (left) towards the present (right). Assuming B and C will give rise to all contemporary biodiversity, the Last Universal Common Ancestor (LUCA) must have come into existence at timepoint t1. **[RJCB: TODO: Use proper phylogeny with proper timescale]**

The biodiversity derived from LUCA is important to us humans (apart from that is has created us) for many reasons. Biodiversity is found so important, that, for example, the European Union has an explicit Biodiversity Strategy, which aims to halt the loss of biodiversity and ecosystem services (`https://ec.europa.eu/environment/nature/biodiversity/strategy/index_en.htm`). Ecosystem services are features of biological systems that are positive for human well-being, for example food, carbon sequestration, waste decomposition and pest control. A review paper **?** shows that biodiversity unsually improves ecosystem services.

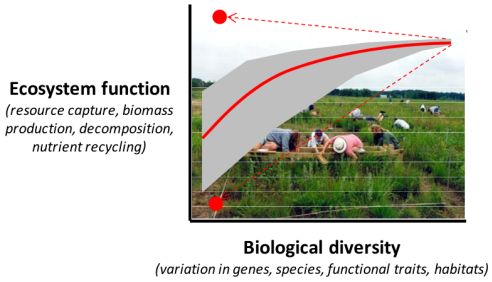

Figure 1.4 | A diversity-function relationship found to be typical from hundreds of studies. The red line represents an average, where the grey polygon represents a 95% confidence interval. The red dots show the lower and upper limit for monocultures. From **?**

Speciation is the process that increases biological diversity. This process is studied from multiple angles, among others, we can study the mechanism ('what causes a speciation event?') or we can study the patterns of many of such events ('is speciation rate constant through time?')

The mechanism of a speciation event has many facets. For more than half a century ago, it was hypothesized that speciation is caused by geographical isolation (e.g. Mayr, 1942) or due to ecological factors (e.g. Lack, 1947). [other proximate causes of speciation]

Instead of looking at the mechanism behind each speciation event, we can also look

at patterns of speciation events through evolutionary time, asking question such as 'How often do speciation and extinction events take place?' 'Are speciation and extinctions rates constant or do they change?', 'What causes a change in speciation or extinction rate?' or 'Is there an upper limit on the number of species?'.

There are two methods to research speciation patterns back in evolutionary time: the use of fossils or using molecular phylogenies.

Using fossils is a classic way to look back in evolutionary time. Fossils show a glimpse of the biodiversity in the past. We can deduce the age of fossils, by the rock layers they are found in. Using fossils has its limitations. First, it is mostly species with hard body parts that are suitable to fossilize. Of such species, an organisms is still only rarely preserved, of which only a fraction under ideal circumstances. Of these fossils, only a fraction is discovered. One example of a famous fossil is 'El Graeco', which may be the oldest known hominin **?**, where homonins are the tribe (taxanomic group) we Homo sapiens share with the Panini.

Using molecular phylogenies is the modern way to look back in evolutionary time. It is the use of heritable molecules (e.g. DNA, RNA, or protein) of contemporary species to infer phylogenies. The field of phylogenetics is the research discipline that intends to infer the most accurate phylogenies possible, regarding topology, speciation and extinction times, optionally adding morphological data and/or fossil data. Phylogenetics is applied in many settings, among others, species classification, forensics, conservation ecology and epidemiology **?**.

One example of the importance of an accurate phylogenetic tree is demonstrated in **?**. This study investigated which loci of the H3 hemagglutinin surface protein are under selection, by constrasting asynonymous and synonymous mutation rates along the branches of a phylogeny. In a preliminary analysis by the authors, they noted that most selection rates were either below or above the statistical threshold depending on the phylogeny. This study contributed to the selection of recommended composition of influenza virus vaccines.

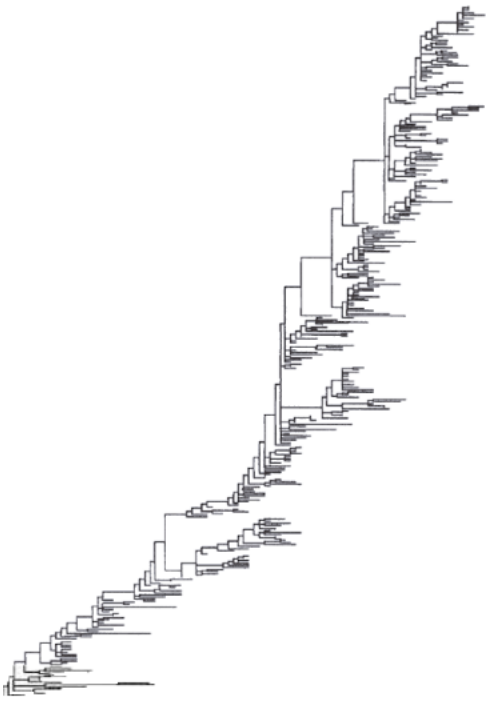

Figure 1.5 | MP tree

Another example of the importance of an accurate phylogenetic tree comes from conservation biology, in which phylogenies are used to calculate an EDGE ('Evolutionarily Distinct and Globally Endangered') score. Species with a high EDGE score are prioritized in conservation. To calculate an EDGE score, one needs a metric of evolutionary distinctiveness ('ED') and globally 'endangeredness' ('GE'). The GE score is a conservational status, ranging from from zero ('Least Concern') to four ('Critically Endangered'). The ED embodies the amount of evolutionary history lost when a species would go extinct, which can be calculated from a (hopefully accurate) phylogeny.

Phylogenetics has taken a huge flight, due to the massively increased computational power and techniques. A first milestone in this file is Felsensteyn's work in 1980, creating (and still maintaining) PHYLIP, the first software package for classical phylogenetic analysis. Another milestone is the Metropolis-Hasting algorithm, which allowed Bayesian phylogenetics to thrive, resulting in contemporay tools such as BEAST, BEAST2, MrBayes and RevBayes.

A clear example of the power of modern phylogenetics, is the Tree Of Life: it uses 3,083 genomes of 2,596 amino-acid positions to create one big phylogeny of all (sequenced) life on Earth, which took 3,840 computational hours on a modern supercomputer [Hug et al., 2016].

To create such a tree from protein sequences, one has to specify an evolutionary model. This evolutionary model embodies our set of assumptions, such as the evolution of a protein sequence (also called the site model), the rate(s) at which this happens (the

clock model) and the rate(s) at which a branching/speciation event takes place (the tree model). For example, the amino acids of the Tree Of Life are assumed to change/mutate according to the LG [Si Quang Le and Olivier Gascuel, An Improved General Amino Acid Replacement Matrix] model, which is a model that uses the average transition rates found in nature.

There are many evolutionary models to choose from, and selecting which one to use is hard, due to the many sets of assumptions to choose from. In general, modellers are looking for that set of assumptions that is as simple as possible, but not simpler. And even then, sometimes an overly simplistic model is picked regardless, due to computation constraints.

By using a model comparison, one has a rational way to select an evolutionary model that is as simple as possible, but not simpler. A model comparison algorithm selects the evolutionary model that is most likely to have generated the data, without being overly complex. The idea is that the best evolutionary model should result in the most accurate phylogenetic trees.

Because model comparison is hard, there have been multiple studies that investigate the effect of picking the wrong evolutionary models.

One example that demonstrates the effect of using a too simple inference model comes from Revell at al **?**. They simulated many phylogenies and respective DNA sequences using different DNA substitution models. After this, they inferred phylogenies from the simulated alignments with either the correct or a simpler DNA substitution model. Ideally, the inferred phylogenies match the phylogenies the alignments are based upon. They found that when the DNA model is the correct one, inference of the phylogenies is satisfactory. However, when using an overly simplistic DNA model, the inferred trees show a slowdown in their speciation rates, also when the original trees were simulated with a constant speciation rate. This study shows that a decreasing speciation rate may be attributed to an overly simplistic DNA model, instead of an interesting biological process.

A more recent example that demonstrates the effect of using a too simple inference model is about assuming a wrong clock model. A clock model embodies our assumptions regarding the mutation rates in the history of different taxa. The simplest clock model, called the strict clock model, assumes these mutation rates are equal across all taxa. Using a wrong clock model has a profound impact on the inferred phylogenetic trees, unless we can specify the timing of some early speciation events **?**.

The tree model is the most important piece of an evolutionary model, with regard to speciation. The assumptions of a tree model is called the tree prior, where 'prior' refers to the knowledge known before creating a phylogeny. The tree prior specifies how likely processes that determine the shape of a tree occur. These two processes are the formation of a new branch and the termination of an existing branches. In the context of speciation, we call these two events a speciation and an extinction event respectively.

There are two standard tree models, called the Yule (also: pure-birth(and (standard) birth-death model. The most basic speciation model is the Yule model [Yule, 19..] which assumes that speciation is constant and there is no extinction. [Research on fossils with Yule model would be fun]. The Yule model predict that the number of extant species grows exponentially through time.

The Birth-Death model [Nee et al., 1994] is an extension of the Yule that allows for a

constant extinction rate. If the speciation rate exceeds the extinction rate, also the BD model predicts that the number of extant species grows exponentially through time. If the extinction rate exceeds the speciation rate, the number of lineages is expected to decline exponentially. The latter is biologically irrelevant.

It is clear that an exponential growth in the expected number of lineages is biologically nonsense. To state the obvious: a finite area (Earth) results in a finite number of species. Applying the BD model to molecular data already shows that it does not always hold, see figure [below]

A recent study that investigates the effect of picking a wrong standard tree prior, comes from Sarver et al, 2019 **?**. In this study, they first simulate trees using either a Yule or a birth-death tree model, after which they simulate an alignment from that phylogeny using two different standard clock models. From these alignments, they inferred the original trees using all of the four different clock and tree prior combinations. They show that, regardless which priors are used, the estimated speciation and diversification rates from the inferred trees are similar to those of the original tree.

This thesis investigates the effect of picking a wrong standard tree prior, when the tree is generated by a non-standard, novel tree model. I will described the two new biological tree models that have been investigated, as well as the re-usable framework to do so.

The first novel and non-standard tree model is the protracted birth-death model (PBD) [Etienne and Rosindell, 201?]. Where the standard BD models assume that a speciation event creates two new species instantly, the PBD model assumes that one of these two species is an incipient species. An incipient species is a new species that is not yet recognized as such, although complete reproductive isolation is already present. A biological example is from [Fennesy, 201?] in which some new giraffe species have been discovered by sequencing part of their DNA. Although these new species have been 'discovered' recently, the had been no gene flow between species for already two million years.

Using the BD model in species that are slow to speciate, will cause an underestimation of the number of lineages in the present (as in the giraffes), in effect possibly giving the illusion that speciation slows down, where in reality it does not.

The second novel and non-standard tree model is the multiple-birth death (MBD) model [Laudanno et al., 2020]. Where the standard BD models assume that a speciation event occurs in one species only at a time, the MBD models allows for speciation events to occur in multiple species at the same time. The biological idea behind this model, is that when a habitat (lake or mountain range) gets split into two, this may trigger speciation events in multiple species of both communities at the same time. This mechanism is posed as an explanation for high biodiversity in lake Tangyanika, where the water level rises and falls with ice ages, splitting up and merging the lake again and again, triggering co-occurring speciation events each change.

This thesis investigates the effect of picking a wrong standard tree prior, when the tree is generated by the non-standard PBD (chapter 5) or MBD (chapter 4) tree model. It does so, by using the same experimental setup, called 'pirouette', which is described in chapter 3. This framework is built up a foundation of R packages called 'babette', which is described in chapter 2.

In the end, we want to know how well we can infer a phylogeny from molecular data

found in the field. That field, outside, which follows an unknown speciation model. Let us just hope our inference is robust to whatever novel model we throw at it.

1.1. PHOTO ATTRIBUTION

Figure **??**, 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 **??**, 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.