Phylogenetic Biology

Casey W. Dunn

2020-07-26

Contents

Preface

I developed this book as a collection of the concepts, methods, and applications that I most wanted to share with students in my Phylogenetic Biology courses.

Chapter 1

Introduction

Phylogenetic biology is the study of evolutionary relationships.

There are three components to a phylogenetic analysis: characters, evolutionary models, and phylogenetic trees. In any given study we could have data, not care, or want to infer any three of these.

When we don't care, we marginalize. When we have data we fix. When we want to infer we estimate.

This gives a broad range of hypothetical phylogenetic investigations. Only a subset of these are common in the literature, due to the questions that drive the field right now, the available data, and the methods that have been implemented in software tools.

When we want to infer phylogenetic trees, we usually fix the characters at the tips using our data, marginalize over historical character states, estimate models, and estimate the tree.

When we want to infer the history of character evolution, we often fix the tree, fix the characters at the tips using our data, estimate the model, and estimate the character history.

But we can also marginalize over models, for example.

1.1 Why phylogenies matter

History of life on earth

Phylogenetic systematics

Comparative methods

Processes of evolution

Current and future evolution (eg epidemiology)

1.2 Required background

Comfort at the command line, manipulation of text files. *Practical Computing for Biologists.*

Basic literacy with the programming language R. *R for Data Science.*

Basic statistics, in particular probability theory. Some math, in particular an understanding of matrices and manipulations of matrices.

1.3 Additional resources

Chapter 2

Trees

2.1 Anatomy

Trees are a special find of graph. Bifurcating acyclic.

Nodes, edges (branches).

I'll stick with the mathematical convention of calling branches edges, since this makes for stronger conceptual ties to other fields that also consider graphs.

2.2 Types

Cladogram, phylogram, time tree

2.3 Interpretation

Tree thinking

2.4 Representation

paranthetical

ape

ggtree

Chapter 3

Simulation

Build the machinery to simulate the evolution of traits on trees. Focus for now on DNA evolution.

3.1 Models

Generative models

The intent is a simplified representation of the process under consideration.

“All models are wrong, some are useful”

3.2 A simple model

Let’s start with a simple model of DNA evolution. Imagine that when the DNA is being replicated, most of the time the appropriate nucleotide is incorporated. Some fraction of the time, at rate μ, an event occurs where the appropriate nucleotides is replaced with a random nucleotide instead. The probability of selecting any of the nucleotides during one of these random replacement events is uniform (picking a C is just as probably as picking a G, for example), and the new nucleotide doesn’t depend in any way on what nucleotide was there before. It is as if you had a bag containing equal frequencies of C, G, T, and A nucleotides. As you built the new DNA strand, every so often you would replace the nucleotide you should be adding with one you instead selected by reaching into the bag with your eyes closed and picking one at random.

Not all replacement events will result in an apparent change. Sometimes the appropriate nucleotide is selected by chance, even though it was picked at random. If, for example, the appropriate nucleotide was an A, under this model

1/4 of the time a replacement event occurs an A is selected by chance and there is no apparent change. In such a case, there has not been a substitution (just a replacement in kind). If the A is replaced with any of the other three nucleotides we say there has been a substitution. Because three of the four possible outcomes of an event result in a substitution, the substitution rate is 3β, which, because $\beta = \mu/4$, is equivalent to noting that the substitution is $(3/4)\mu$. Because some events result in no apparent change, substitutions are only a subset of events and the substitution rate is *lower* than the replacement event rate.

It might seem a bit odd to consider replacement events that don't result in substitutions, but this follows naturally from a central feature we specified for the the model - the new nucleotide doesn't depend in any way on what nucleotide was there before. If we had a process where replacements always resulted in substitutions, then excluding the a replacement in kind would require knowing which nucleotide should be placed so that we *don't* select it.

3.2.1 Expected amount of change

One of the primary values of a model is that it allows us to think explicitly about how much evolutionary change we expect to see under the specified process. For the simple process described here, there are two things to consider if we want to know the amount of evolutionary change. The first is the substitution rate μ (which we also know if we know β, since $\mu = 4\beta$), and the time over which the evolutionary process acts.

In Figure 3.1 the amount of evolutionary time is held constant, and the rate μ is changed. When $\mu = 0$, the bottom bar, there are no replacements (black bars) and therefore no substitutions (the whole bar is the same color).

As $\mu = 0$ increases (going up on the y axis), the number of replacement events over the same time interval increases (Figure 3.2). This reflects the simple linear relationship $n = \mu t$, where n is the number of expected replacement events.

Because of the linear relationship between the number of replacements and the product μt, rate and time are conflated. In many scenarios you can't estimate them independently. If there are a small number of replacements, for example, you can't be sure if there is a low rate over a long time interval or a high rate over a short interval. Both would give the same result. Because they are so often confounded in phylogenetic questions, often the rate is essentially fixed at one and the unit of time for edge lengths is given as the number of expected evolutionary change rather than absolute time (years, months, etc). You will often see this length as the scale bar of published phylogenies (Figure 3.3). The exception is when you have external information, such as dated fossils, that allow you to independently estimate edge lengths and rates.

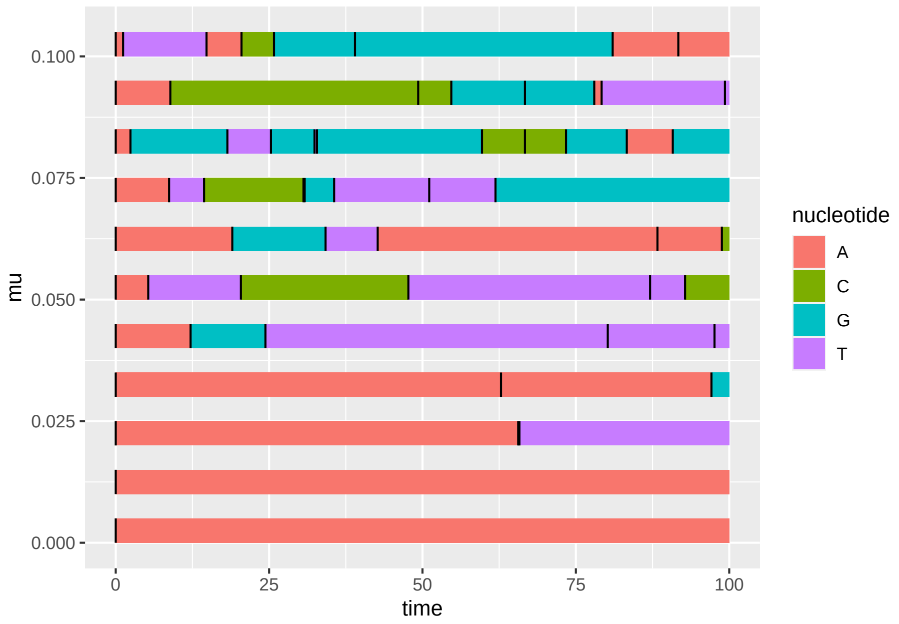

Figure 3.1: Each horizontal bar is a simulation of evolution of a single nucleotide position through time, t, for a specified value of μ. Each sumulation starts out as an A. Black vertical bars correspond to replacement events, which don't all lead to substitutions (a new color).

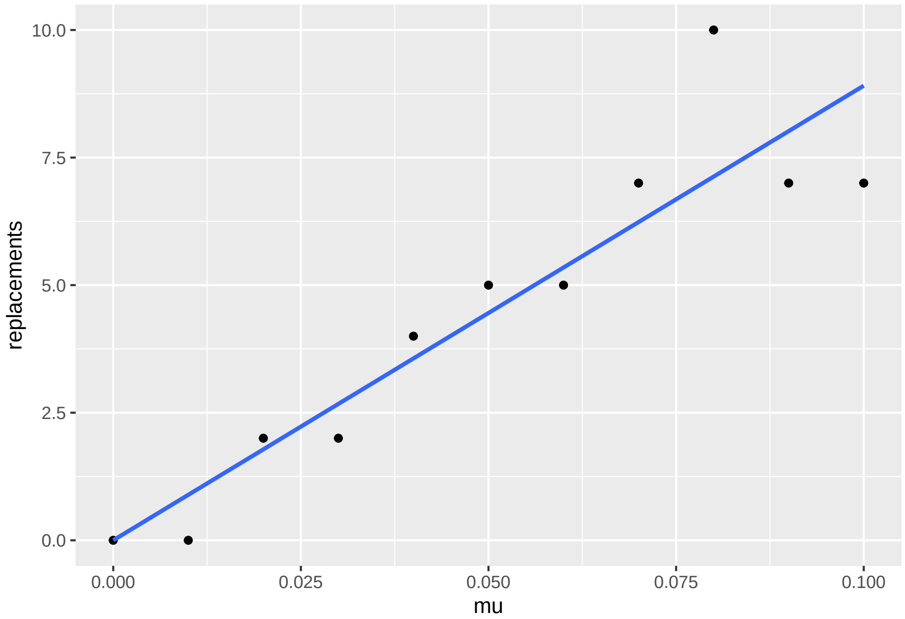

Figure 3.2: The number of replacement events increases linearly with the replacement rate μ. This plot is from the same simulation as that shown in Figure 3.1. The line is a linear model fit to the data.

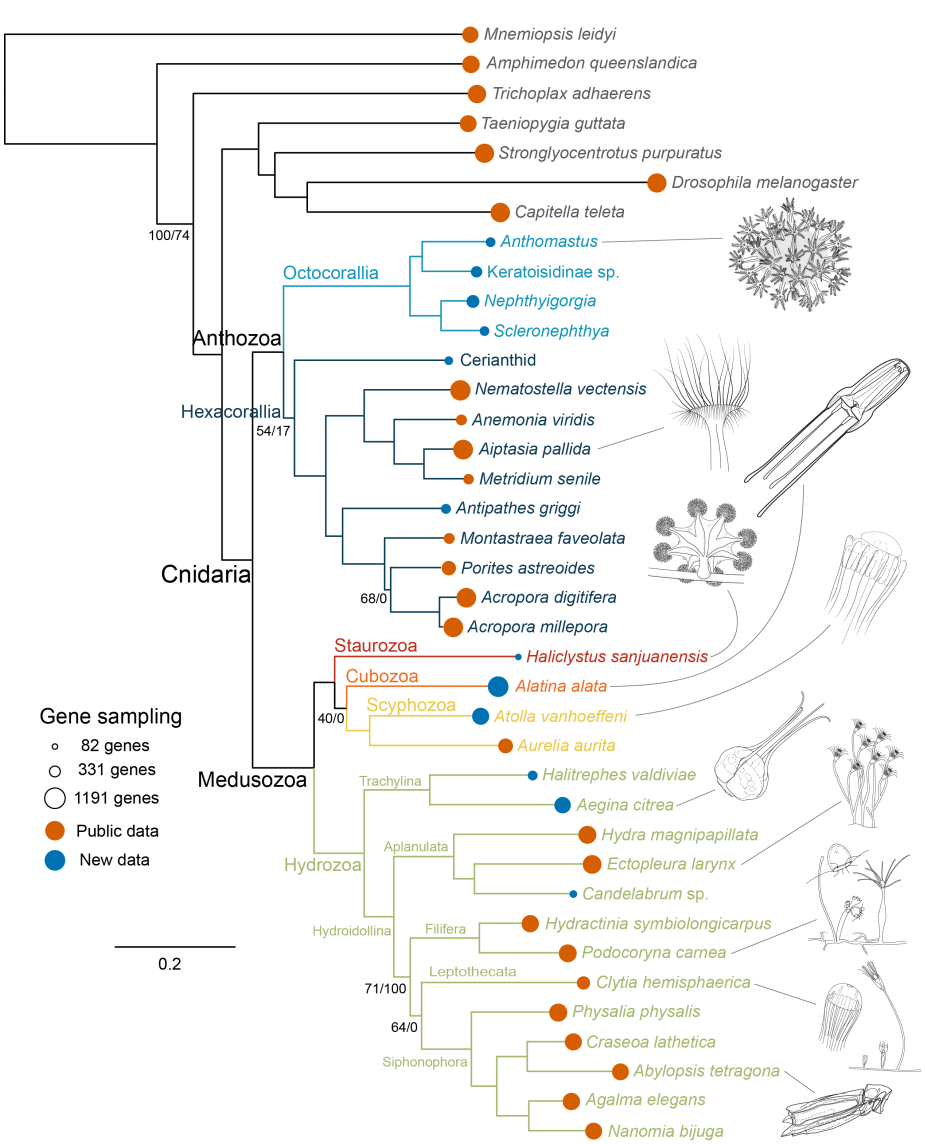

Figure 3.3: A published phylogeny (Zapata et al., 2015) with a scale bar indicating branch length in terms of the expected amount of evolutionary change, rather than absolute time.

3.2.2 Expected end state

The machinery above shows how a model can clarify the way we think about the expected amount of change. Many times, though, we want to know what the probability of a given end state is given a starting state, a model, and the amount of time elapsed. One way to anchor such a question is to think about the extremes - what do we expect after a very small amount of change (either a short time or a slow rate of change, or both), and what do we expect after a large amount of change?

The situation is most clear after a small amount of change - we expect the end result to be the same as the starting condition. If we start with an A, for example, we expect to end with an A. In this situation, if we know the starting state that information tells us a lot about the end state. Not much else matters.

What should we expect, though, if there has been a large amount of change? Can we know anything at all? It turns out that we can. If there have been many replacements, one after the other, than the initial starting state really doesn't matter at all because whatever was there will probably have been replaced multiple times. If the starting state doesn't contain information about the end state, what does? It is the bag that you are picking the nucleotides at random from. Given enough evolutionary time, our simple model will lead the expected frequency of each nucleotide in the evolving sequence to be the same as the frequency in the bag that we randomly draw them from. Since we specified that you have the same chance of grabbing any nucleotide from the bag, eventually the probability of having each of the our nucleotides is the same, 25%. If you started with a sequence that had an A and let it evolve 100 times, after enough evolutionary time had passed to reach equilibrium you would expect to get 25 C's, 25 G's, 25 T's, and 25 A's.

3.3 Generalizing the simple model

rates, equilibrium frequencies

```
##      [,1]
## [1,]    0
## [2,]    0
## [3,]    0
## [4,]    0
```

exponentiation

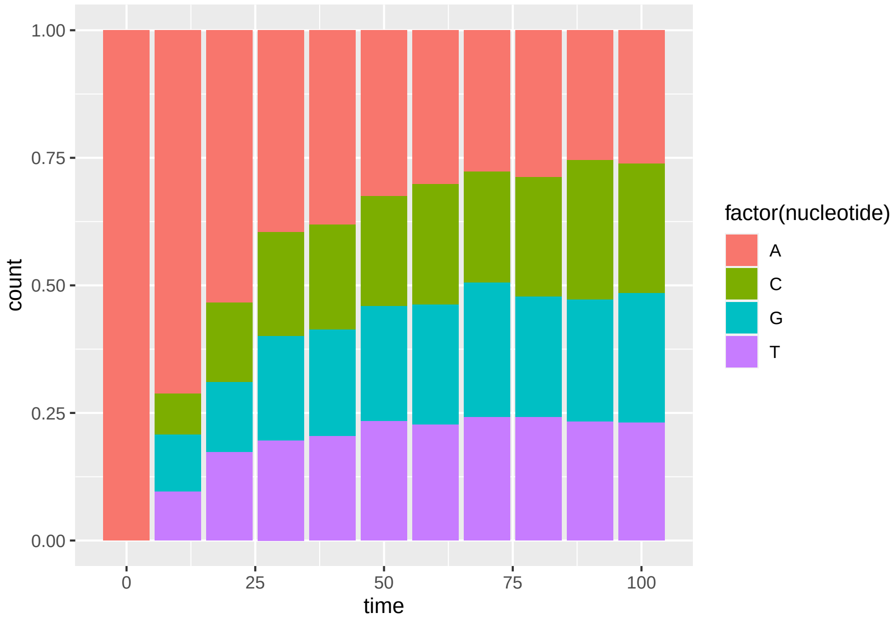

Figure 3.4: Stacked bar plots indicating the frequency of each nucleotide after evolution for a specified amount of time. The rate of evolution is \$ = 0.05 \$. There are 1000 replicate simulations for each value of time. At time=0 (no evolution), the end result is always the same as the initial value, which is fixed at A in these simulations. As the length of time increases, the four nucleotides converge on equal frequencies of 25% each.

3.4 More complex models

3.5 Model structure

3.6 Aditional resources

- My own thinking about this material was heavilly influenced by Paul Lewis's wonderful lectures at the annual Workshop on Molecular Evolution at Woods Hole. Some of his lectures are now available online as part of the excellent Phylo Seminar, starting with https://www.youtube.com/watch?v=1r4z0YJq580&t=2111s

Chapter 4

Inference

A specific task - given a set of character data corresponding to the tips of a tree, what is the topology of the tree? Model is also estimated, but may or may not be of interest.

Calculating the likelihood of a tree

What a likelihood is

maximum likelihood

hueristics

Bayesian

“Model free” methods

Chapter 5

Evaluation

5.1 Model evaluation

Why not add as many parameters as you can imagine?

Likelihood ratio test

AIC

BIC

5.2 Topology evaluation

5.3 Sensitiviy

changing methods and parameters

adding noise

5.4 Confidence

Propogation, point estimates

5.5 Epistemology

gene trees, species trees

Chapter 6

Future

Integrated models of genome evolution

Chapter 7

Shape

Dating

Diversification

Extinction, birth, death

Null models

Chapter 8

Character evolution

Comparative biology

Reconstructing a single trait on trees

Trait correlation

Models of character change, including rate

Chapter 9

Applications

9.1 Functional genomic data

9.2 Gene trees

Bibliography

Zapata, F., Goetz, F. E., Smith, S. A., Howison, M., Siebert, S., Church, S. H., Sanders, S. M., Ames, C. L., McFadden, C. S., France, S. C., et al. (2015). Phylogenomic analyses support traditional relationships within cnidaria. *PloS one*, 10(10):e0139068.