pdf version

Chapter 1: A Macroevolutionary Research Program

Section 1.1: Introduction

Evolution is happening all around us. In many cases – lately, due to technological advances in molecular biology – scientists can now describe the evolutionary process in exquisite detail. For example, we know exactly which genes change in frequency from one generation to the next as mice and lizards evolve a white color to match the pale sands of their novel habitats (Rosenblum et al. 2010). We understand the genetics, development, and biomechanics processes that link changes in a Galapagos finches' diet to the shape of their bill (Abzhanov et al. 2004). And, in some cases, we can watch as one species splits to become two (for example, Rolshausen et al. 2009).

Detailed studies of evolution over short time-scales have been incredibly fruitful and important for our understanding of biology. But evolutionary biologists have always wanted more than this. Evolution strikes a chord in society because it aims to tell us how we, along with all the other living things that we know about, came to be. This story stretches back some 4 billion years in time. It includes all of the drama that one could possibly want – sex, death, great blooms of life and global catastrophes. It has had "winners" and "losers," groups that wildly diversified, others that expanded then crashed to extinction, as well as species that have hung on in basically the same form for hundreds of millions of years.

There is, perhaps, no more evocative symbol of this grand view of evolution over deep time than the tree of life (Figure 1.1). This branching phylogenetic tree connects all living things through a series of splitting branches to a single common ancestor. Recent research has dramatically increased our knowledge of the shape and form of this tree. The tree of life is a rich treasure-trove of information, telling us how species are related to one another, which groups are exceptionally diverse or depauperate, and how life has evolved, formed new species, and spread over the globe. Our understanding of the tree of life, still incomplete but advancing every day, promises to transform our understand of evolution at the grandest scale.

Knowing the evolutionary processes that operate over the course of a few generations, even in great detail, does not automatically give insight into why the tree of life is shaped the way that it is. At the same time, it seems reasonable to hypothesize that the same processes that we can observe now - natural selection, genetic drift, migration, sexual selection, and so on - have been occurring for

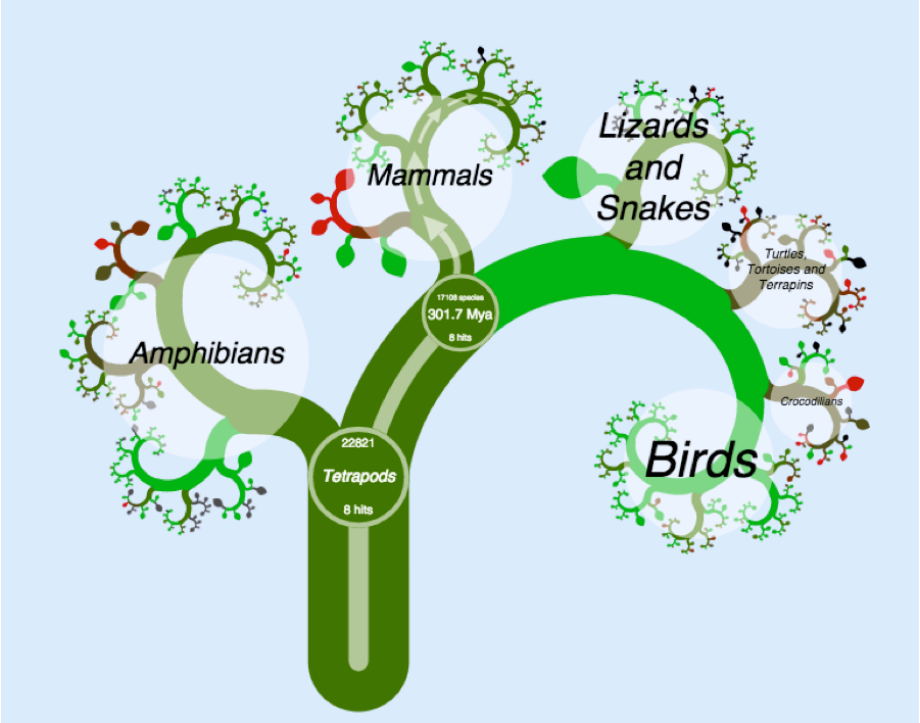

Figure 1: Figure 1.1. A small section of the tree of life showing the relationships among tetrapods, from onezoom (Rosindell and Harmon 2012). Arrows lead to you.

the last four billion years or so along the branches of the tree. A major challenge for evolutionary biology, then, comes in connecting our knowledge of the mechanisms of evolution with broad-scale patterns seen in the tree of life. This "tree thinking" is what we will explore.

In this book, I describe methods to connect evolutionary processes to broad-scale patterns in the tree of life. I focus mainly – but not exclusively – on phylogenetic comparative methods. Comparative methods combine biology, mathematics, and computer science to learn about a wide variety of topics in evolution (see Harvey and Pagel 1991 for an early review). For example, we can find out which processes must have been common, and which rare, across clades in the tree of life; whether evolution has proceeded differently in some lineages compared to others; and whether the evolutionary potential that we see playing out in real time is sufficient to explain the diversity of life on earth, or whether we might need additional processes (like adaptive radiation or species selection) that may come into play only very rarely or over very long timescales.

This introductory chapter has three sections. First, I lay out the background and context for this book, highlighting the role that I hope it will play for readers. Second, I include some background material on phylogenies - both what they are, and how they are constructed. This is necessary information that leads into the methods presented in the remainder of the chapters of the book; interested readers can also read Felsenstein (Felsenstein 2004), which includes much more detail. Finally, I briefly outline the book's remaining chapters.

Section 1.2: The roots of comparative methods

The comparative approaches in this book stem from and bring together three main fields: population and quantitative genetics, paleontology, and phylogenetics. I will provide a very brief discussion of how these three fields motivate the models and hypotheses in this book (see Pennell and Harmon 2013 for a more comprehensive review).

Population and quantitative genetics models quantify how gene frequencies and trait values change through time. These models lie at the core of evolutionary biology, and relate closely to a number of approaches in comparative methods. Population genetics tends to focus on allele frequencies, while quantitative genetics focuses on traits and their heritability; however, genomics has begun to blur this distinction a bit. Both population and quantitative genetics approaches have their roots in the modern synthesis, especially the work of Fisher (1930) and Wright (1984), but both have been greatly elaborated since then (see Lynch and Walsh 1998; Rice 2004). Although population and quantitative genetic approaches most commonly focus on change over one or a few generations, they have been applied to macroevolution with great benefit. For example, Lande (1976) provided quantitative genetic predictions for trait evolution over many generations using Brownian motion and Ornstein-Uhlenbeck models (see Chap-

ter 3). Lynch (1990) later showed that these models predict long-term rates of evolution that are actually too fast; that is, variation among species is too small compared to what we know about the potential of selection and drift to change traits. This is, by the way, a great example of the importance of macroevolutionary research from a deep-time perspective. Given the regular observation of strong selection in natural populations, who would have guessed that long-term patterns of divergence are actually less than we would expect, even considering only neutral genetic drift alone (see also Uyeda et al. 2011)?

Paleontology has, for obvious reasons, focused on macroevolutionary models as an explanation for the distribution of species and traits in the fossil record. Almost all of the key questions that I tackle in this book are also of primary interest to paleontologists. For example, a surprising number of the macroevolutionary models and concepts in use today stem from quantitative approaches to paleobiology by Raup and colleagues in the 1970s and 1980s (e.g. Raup et al. 1973; Raup 1985). Many of the models that I will use in this book – for example, birth-death models for the formation and extinction of species – were first applied to macroevolution by paleobiologists.

Finally, comparative methods has deep roots in phylogenetics. In fact, many modern phylogenetic approaches to macroevolution can be traced to Felsenstein's (1985) paper introducing independent contrasts. This paper was unique in three main ways. First, Felsenstein's paper was written in a remarkably clear way, and convinced scientists from a range of disciplines of the necessity and value of placing their comparative work in a phylogenetic context. Second, the method of phylogenetic independent contrasts was computationally fast and straightforward to interpret. And finally, Felsenstein's work suggested a way to connect the previous two topics, quantitative genetics and paleobiology, using math. I discuss independent contrasts, which continue to find new applications, in great detail later in the book. Felsenstein (1985) spawned a whole industry of quantitative approaches that apply models from population and quantitative genetics, paleobiology, and ecology to data that includes a phylogenetic tree.

25 years ago, "The Comparative Method in Evolutionary Biology," by Harvey and Pagel (1991) synthesized the new field of comparative methods into a single coherent framework. Even reading this book nearly 25 years later one can still feel the excitement and potential unlocked by a suite of new methods that use phylogenetic trees to understand macroevolution. But in the time since Harvey and Pagel (1991), the field of comparative methods has exploded – especially in the past decade. Much of this progress was, I think, directly inspired by Harvey and Pagel's book, which went beyond review and advocated a model-based approach for comparative biology. My wildest hope is that our own book can serve a similar purpose.

My goals in writing this book, then, are three-fold. First, to provide a general introduction to the mathematical models and statistical approaches that form the core of comparative methods; second, to give just enough detail on statistical machinery to help biologists understand how to tailor comparative methods

to their particular questions of interest, and to help biologists get started in developing their own new methods; and finally, to suggest some ideas for how comparative methods might progress over the next few years.

Section 1.3: A brief introduction to phylogenetic trees

It is hard work to reconstruct a phylogenetic tree. This point has been made many times (for example, see Felsenstein 2004), but bears repeating here. There are an enormous number of ways to connect a set of species by a phylogenetic tree – and the number of possible trees grows extremely quickly with the number of species. For example, there are about 5 x 1038 ways to build a phylogenetic tree* of 30 species, which is many times larger than the number of stars in the universe. Additionally, the mathematical problem of reconstructing trees in an optimal way from species' traits is an example of a problem that is "NP-complete," a class of problems that include some of the most computationally difficult in the world. Building phylogenies is difficult.

The difficulty of building phylogenies is currently reflected in the challenge of reconstructing the tree of life. Some parts of the tree of life are still unresolved even with the tremendous amounts of genomic data that are now available. Accordingly, scientists have devoted a focused effort to solving this difficult problem. There are now a large number of fast and efficient computer programs aimed solely at reconstructing phylogenetic trees (e.g. MrBayes: Ronquist and Huelsenbeck 2003; BEAST: Drummond and Rambaut 2007). Consequently, the number of well-resolved phylogenetic trees available is also increasing rapidly. As we begin to fill in the gaps of the tree of life, we are developing a much clearer idea of the patterns of evolution that have happened over the past 4.5 billion years on Earth.

The core reason that phylogenetic trees are difficult to reconstruct is that they are information-rich. A single tree contains detailed information about the patterns and timing of evolutionary branching events through a group's history. Each branch in a tree tells us about common ancestry of a clade of species, and the start time, end time, and branch length tell us about the timing of speciation events in the past. If we combine a phylogenetic tree with some trait data – for example, mean body size for each species in a genus of mammals – then we can obtain even more information about the evolutionary history of a section of the tree of life.

The most common methods for reconstructing phylogenetic trees use data on species' genes and/or traits. The core information about phylogenetic relatedness of species is carried in shared derived characters; that is, characters that have evolved new states that are shared among all of the species in a clade and not found in the close relatives of that clade. For example, mammals have many shared derived characters, including hair, mammary glands, and specialized inner ear bones.

Phylogenetic trees are often constructed based on genetic (or genomic) data using modern computer algorithms. Several methods can be used to build trees, like parsimony, maximum likelihood, and Bayesian analyses (see chapter 2). These methods all have distinct assumptions and can give different results. In fact, even within a given statistical framework, different software packages (e.g. Mr. Bayes and BEAST, both Bayesian approaches) can give different results for phylogenetic analyses of the same data. The details of phylogenetic tree reconstruction are beyond the scope of this book. Interested readers can read "Inferring Phylogenies" (Felsenstein 2004), "Computational Molecular Evolution" (Yang 2006), or other sources for more information.

For many current comparative methods, we take a phylogenetic tree for a group of species as a given – that is, we assume that the tree is known without error. This assumption is almost never justified. There are many reasons why phylogenetic trees are estimated with error. For example, estimating branch lengths from a few genes is difficult, and the branch lengths that we estimate should be viewed as uncertain. As another example, trees that show the relationships among genes (gene trees) are not always the same as trees that show the relationships among species (species trees). Because of this, the best comparative methods recognize that phylogenetic trees are always estimated with some amount of uncertainty, both in terms of topology and branch lengths, and incorporate that uncertainty into the analysis. I will describe some methods to accomplish this in later chapters.

How do we make sense of the massive amounts of information contained in large phylogenetic trees? The definition of "large" can vary, but we already have trees with tens of thousands of tips, and I think we can anticipate trees with millions of tips in the very near future. These trees are too large to comfortably fit into a human brain. Current tricks for dealing with trees – like banks of computer monitors or long, taped-together printouts – are inefficient and will not work for the huge phylogenetic trees of the future. We need techniques that will allow us to take large phylogenetic trees and extract useful information from them. This information includes, but is not limited to, estimating rates of speciation, extinction, and trait evolution; testing hypotheses about the mode of evolution in a group; identifying adaptive radiations, key innovations, and other macroevolutionary explanations for diversity; and many other things.

Section 1.4: What we can (and can't) learn about evolutionary history from living species

Traditionally, scientists have used fossils to quantify rates and patterns of evolution through long periods of time (sometimes called "macroevolution"). These approaches have been tremendously informative. We now have a detailed picture of the evolutionary dynamics of many groups, from hominids to crocodilians. In some cases, very detailed fossil records of some types of organisms – for example,

marine invertebrates – have allowed quantitative tests of particular evolutionary models.

Fossils are particularly good at showing how species diversity and morphological characters change through time. For example, if one has a sequence of fossils with known times of occurrence, one can reconstruct patterns of species diversity through time. A classic example of this is Sepkoski's (1984) reconstruction of the diversity of marine invertebrates over the past 600 million years. One can also quantify the traits of those fossils and measure how they change across various time intervals (e.g. Foote 1997). In some groups, we can make plots of changes in lineage and trait diversity simultaneously (Figure 1.2). Fossils are the only evidence we have for evolutionary lineages that have gone extinct, and they provide valuable direct evidence about evolutionary dynamics in the past.

However, fossil-based approaches face some challenges. The first is that the fossil record is incomplete. This is a well-known phenomenon, identified by Darwin himself (although many new fossils have been found since Darwin's time!). The fossil record is incomplete in some very particular ways that can sometimes hamper our ability to study evolutionary processes using fossils alone. One example is that fossils are rare or absent from some classical examples of adaptive radiation on islands. For example, the entire fossil record of Caribbean anoles, a well-known adaptive radiation of lizards, consists of less than ten specimens preserved in amber (Losos 2009). We similarly lack fossils for other adaptive radiations like African cichlids and Darwin's finches. The absence of fossils in these groups limits our ability to directly study the early stages of adaptive radiation. Another limitation of the fossil record relates to species and speciation. It is very difficult to identify and classify species in the fossil record – even more difficult than it is to do so for living species. It is hard to tell species apart, and particularly difficult to pin down the exact time when new species split off from their close relatives. In fact, most studies of fossil diversity focus on higher taxonomic groups like genera, families, or orders (see, e.g., Sepkoski 1984). These studies have been immensely valuable but it can be difficult to connect these results to what we know about living species. In fact, it is species (and not genera, families, or orders) that form the basic units of evolutionary studies. So, fossils have great value but also suffer from some particular limitations.

Phylogenetic trees represent a rich source of complementary information about the dynamics of species formation through time. Phylogenetic approaches provide a useful complement to fossils because their limitations are very different from the limitations of the fossil record. For example, one can often include all of the living species in a group when creating a phylogenetic tree. Additionally, one can use information from detailed systematic and taxonomic studies to identify species, rather than face the ambiguity inherent when using fossils. Phylogenetic trees provide a distinct source of information about evolutionary change that is complementary to approaches based on fossils. However, phylogenetic trees do not provide all of the answers. In particular, there are certain

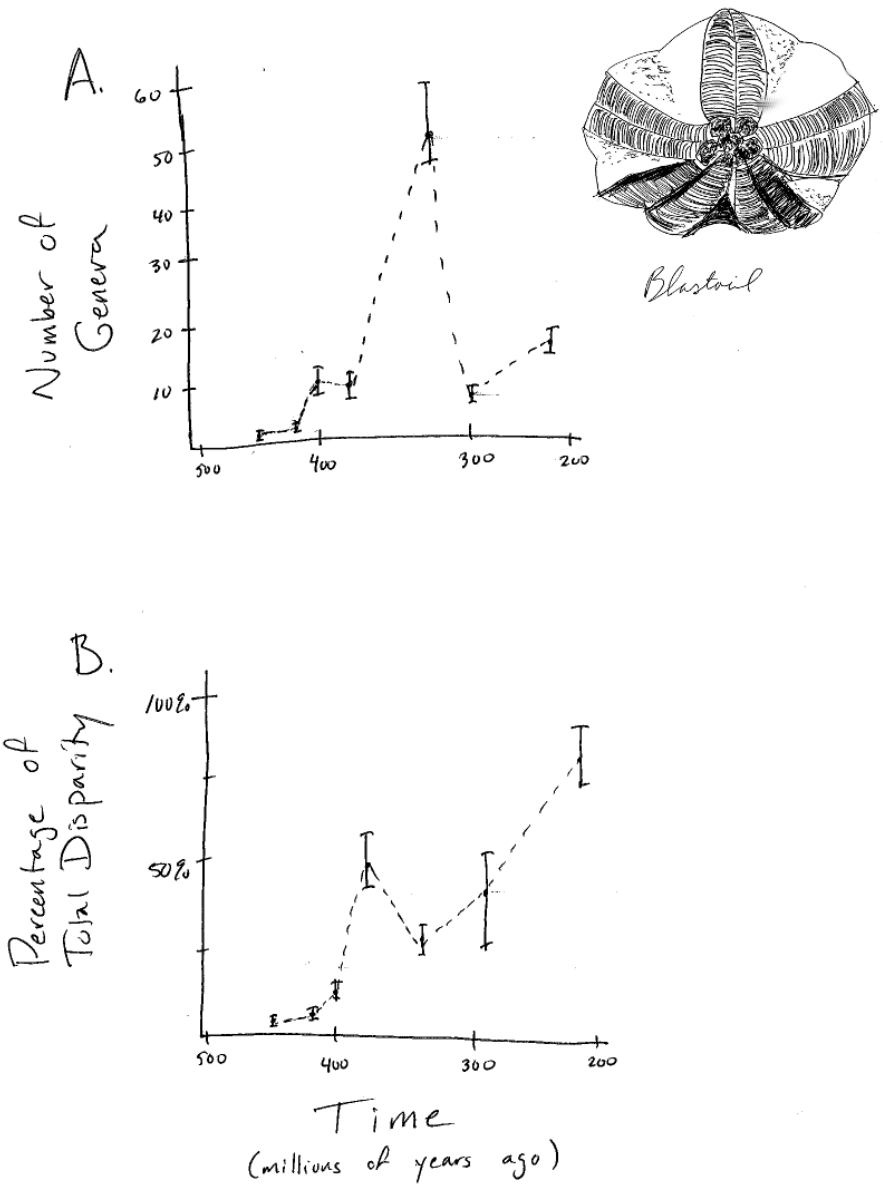

Figure 2: Figure 1.2. Diversity and disparity in the fossil record for the Blastoids. Plots show A. diversity (number of genera) and B. disparity (trait variance) through time. Taken from (Foote 1997).

problems that comparative data alone simply cannot address. The most prominent of these, which I will return to later, are reconstructing traits of particular ancestors (ancestral state reconstruction; Losos 2011) and distinguishing between certain types of models where the tempo of evolution changes through time (Slater et al. 2012). Some authors have argued that extinction, as well, cannot be detected in the shape of a phylogenetic tree (Rabosky 2010) – but I will argue against this point of view in chapter 11. Phylogenetic trees provide a rich source of information about the past, but we should be mindful of their limitations (Alroy 1999).

Perhaps the best approach would combine fossil and phylogenetic data directly. Paleontologists studying fossils and neontologists studying phylogenetic trees share a common set of mathematical models. This means that, at some point, the two fields can merge, and both types of information can be combined to study evolutionary patterns in a cohesive and integrative way. However, surprisingly little work has so far been done in this area (but see Slater et al. 2012).

Section 1.5: Overview of the book

In this book, I outline statistical procedures for analyzing comparative data. Some methods – such as those for estimating patterns of speciation and extinction through time – require an ultrametric phylogenetic tree. Other approaches model trait evolution, and thus require data on the traits of species that are included in the phylogenetic tree. The methods also differ as to whether or not they require the phylogenetic tree to be complete – that is, to include every living species descended from a certain ancestor – or can accommodate a smaller sample of the living species.

The book begins with a general discussion of model-fitting approaches to statistics (Chapter 2), with a particular focus on maximum likelihood and Bayesian approaches. In Chapters 3-9, I describe models of character evolution. I discuss approaches to simulating and analyzing the evolution of these characters on a tree. Chapters 10-12 focus on models of diversification, which describe patterns of speciation and extinction through time. I describe methods that allow us to simulate and fit these models to comparative data. Chapter 13 covers combined analyses of both character evolution and lineage diversification. Finally, in Chapter 14 I discuss what we have learned so far about evolution from these approaches, and what we are likely to learn in the future.

There are a number of computer software tools that can be used to carry out the methods described here. In this book, I focus on the statistical software environment R. For each chapter, my course website, in progress, provides sample R code that can be used to carry out all described analyses. I hope that this R code will allow further development of this language for comparative analyses. However, it is possible to carry out the algorithms we describe using other computer software or programming languages (e.g. Arbor).

Statistical comparative methods represent a promising future for evolutionary studies, especially as our knowledge of the tree of life expands. I hope that the methods described in this book can serve as a Rosetta stone that will help us read the tree of life as it is being built.

Section 1.6: References

Abzhanov, A., M. Protas, B. R. Grant, P. R. Grant, and C. J. Tabin. 2004. Bmp4 and morphological variation of beaks in Darwin's finches. Science 305:1462–1465.

Alroy, J. 1999. The fossil record of North American mammals: Evidence for a Paleocene evolutionary radiation. Syst. Biol. 48:107–118.

Drummond, A. J., and A. Rambaut. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7:214.

Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates, Inc., Sunderland, MA.

Felsenstein, J. 1985. Phylogenies and the comparative method. Am. Nat. 125:1–15.

Fisher, R. A. 1930. The genetical theory of natural selection: A complete variorum edition. Oxford University Press.

Foote, M. 1997. The evolution of morphological diversity. Annu. Rev. Ecol. Syst. 28:129–152.

Harvey, P. H., and M. D. Pagel. 1991. The comparative method in evolutionary biology. Oxford University Press.

Lande, R. 1976. Natural selection and random genetic drift in phenotypic evolution. Evolution 30:314–334.

Losos, J. 2009. Lizards in an evolutionary tree: Ecology and adaptive radiation of anoles. University of California Press.

Losos, J. B. 2011. Seeing the forest for the trees: The limitations of phylogenies in comparative biology. Am. Nat. 177:709–727.

Lynch, M. 1990. The rate of morphological evolution in mammals from the standpoint of the neutral expectation. Am. Nat. 136:727–741.

Lynch, M., and B. Walsh. 1998. Genetics and analysis of quantitative traits. Sinauer Sunderland, MA.

Pennell, M. W., and L. J. Harmon. 2013. An integrative view of phylogenetic comparative methods: Connections to population genetics, community ecology,

and paleobiology. Ann. N. Y. Acad. Sci. 1289:90–105.

Rabosky, D. L. 2010. Extinction rates should not be estimated from molecular phylogenies. Evolution 64:1816–1824.

Raup, D. M. 1985. Mathematical models of cladogenesis. Paleobiology 11:42–52.

Raup, D. M., S. J. Gould, T. J. M. Schopf, and D. S. Simberloff. 1973. Stochastic models of phylogeny and the evolution of diversity. J. Geol. 81:525–542.

Rice, S. H. 2004. Evolutionary theory. Sinauer, Sunderland, MA.

Rolshausen, G., G. Segelbacher, K. A. Hobson, and H. M. Schaefer. 2009. Contemporary evolution of reproductive isolation and phenotypic divergence in sympatry along a migratory divide. Curr. Biol. 19:2097–2101.

Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.

Rosenblum, E. B., H. Römpler, T. Schöneberg, and H. E. Hoekstra. 2010. Molecular and functional basis of phenotypic convergence in white lizards at White Sands. Proc. Natl. Acad. Sci. U. S. A. 107:2113–2117.

Rosindell, J., and L. J. Harmon. 2012. OneZoom: A fractal explorer for the tree of life. PLoS Biol. 10:e1001406.

Sepkoski, J. J. 1984. A kinetic model of phanerozoic taxonomic diversity. III. Post-Paleozoic families and mass extinctions. Paleobiology 10:246–267.

Slater, G. J., L. J. Harmon, and M. E. Alfaro. 2012. Integrating fossils with molecular phylogenies improves inference of trait evolution. Evolution 66:3931–3944.

Uyeda, J. C., T. F. Hansen, S. J. Arnold, and J. Pienaar. 2011. The million-year wait for macroevolutionary bursts. Proc. Natl. Acad. Sci. U. S. A. 108:15908–15913.

Wright, S. 1984. Evolution and the genetics of populations, Volume 1: Genetic and biometric foundations. University of Chicago Press.

Yang, Z. 2006. Computational molecular evolution. Oxford University Press.

pdf version

Chapter 2: Fitting Statistical Models to Data

Section 2.1: Introduction

Evolution is the product of a thousand stories. Individual organisms are born, reproduce, and die. The net result of these individual life stories over broad spans of time is evolution. At first glance, it might seem impossible to model this process over more than one or two generations. And yet scientific progress relies on creating simple models and confronting them with data. How can we evaluate models that consider evolution over millions of generations?

There is a solution: we can rely on the properties of large numbers to create simple models that represent, in broad brushstrokes, the types of changes that take place over evolutionary time. We can then compare these models to data in ways that will allow us to gain insights into evolution.

This book is about constructing and testing mathematical models of evolution. In my view the best comparative approaches have two features. First, the most useful methods emphasize parameter estimation over the use of test statistics and p-values. The best of these methods fit models that we care about and estimate parameters that have a clear interpretation. Increasingly, methods can also recognize and quantify uncertainty in our parameter estimates. Second, some very useful methods involve model selection, the process of using data to objectively select the best model from a set of possibilities. When we use a model selection approach, we take advantage of the fact that patterns in empirical data sets will reject some models as implausible and support the predictions of others. This sort of approach can be a nice way to connect the results of a statistical analysis to a particular biological question.

In this chapter, I will first give a brief overview of standard hypothesis testing in the context of phylogenetic comparative methods. However, standard hypothesis testing can be limited in complex, real-world situations, such as those encountered commonly in comparative biology. I will then review two other statistical approaches, maximum likelihood and Bayesian analysis, that are often more useful for comparative methods. This latter discussion will cover both parameter estimation and model selection.

All of the basic statistical approaches presented here will be applied to evolutionary problems in later chapters. It can be hard to understand abstract statistical concepts without examples. So, throughout this part of the chapter, I will refer back to a simple example.

> A common simple example in statistics involves flipping coins. To fit with the theme of this book, however, I will change this to flipping

a lizard (needless to say, do not try this at home!). Suppose you have a lizard with two sides, "heads" and "tails." You want to flip the lizard to help make decisions in your life. However, you do not know if this is a fair lizard, where the probability of obtaining heads is 0.5, or not. As an experiment, you flip the lizard 100 times, and obtain heads 63 of those times. Thus, 63 heads out of 100 lizard flips is your data; we will use model comparisons to try to see what these data tell us about models of lizard flipping.

Section 2.2: Standard statistical hypothesis testing

Standard hypothesis testing approaches focus almost entirely on rejecting null hypotheses. In the framework (usually referred to as the frequentist approach to statistics) one first defines a null hypothesis that represents your expectation if some process of interest were not occurring. For example, perhaps you are interested in comparing the mean body size of two species of lizards, an anole and a gecko. One null hypothesis would be that the two species do not differ in body size. The alternative, which one can conclude by rejecting that null hypothesis, is that one species is larger than the other. Another example might involve investigating two variables, like body size and leg length, across a set of lizard species (I assume here that you have little interest in organisms other than lizards). Here the null hypothesis would be that there is no relationship between body size and leg length. The alternative hypothesis, which again represents the situation where the phenomenon of interest is actually occurring, is that there is a relationship with body size and leg length. For frequentist approaches, the alternative hypothesis is always the negation of the null hypothesis; as you will see below, other approaches allow one to compare the fit of a set of models without this restriction and choose the best amongst them.

The next step is to define a test statistic, some way of measuring the patterns in the data. In the two examples above, we would consider test statistics that measure the difference in mean body size among our two species of lizards, or the slope of the relationship between body size and leg length. One can then compare the value of this test statistic in the data to the expectation of this test statistic under that null hypothesis. The relationship between the test statistic and its expectation under the null hypothesis is captured by a P-value. The P-value is the probability of obtaining a test statistic at least as extreme as the actual test statistic in the case where the null hypothesis is true. You can think of the P-value as a measure of how probable it is that you would obtain your data in a universe where the null hypothesis is true. In other words, the P-value measures how probable it is under the null hypothesis that you would obtain a test statistic at least as extreme as what you see in the data; conversely, if the P-value is very small, then it is extremely unlikely that your data are compatible with this null hypothesis.

If the test statistic is very different from what one would expect under the

null hypothesis, then the P-value will be small: we are unlikely to obtain the test statistic seen in the data if the null hypothesis were true. In that case, we reject the null hypothesis. By contrast, if that probability is large, then there is nothing "special" about your data, at least from the standpoint of your null hypothesis. The test statistic is within the range expected under the null hypothesis, and we fail to reject that null hypothesis. Note the careful language here – in a standard frequentist framework, you never accept the null hypothesis, you simply fail to reject it.

Getting back to our lizard-flipping example, we can use a frequentist approach and carry out a binomial test, which allows us to test whether a given event with two outcomes has a certain probability of success. In this case, we are interested in testing the null hypothesis that our lizard is a fair flipper; that is, that the probability of heads $p_H = 0.5$. The binomial test uses the number of "successes" (we will use the number of heads, $k = 63$) as a test statistic. We then ask whether this test statistic is either much larger or much smaller than we might expect under our null hypothesis. So, our null hypothesis is that $p_H = 0.5$; our alternative, then, is that p_H takes some other value: $p_H \neq 0.5$.

To carry out the test, we consider the distribution of our test statistic (the number of heads) under our null hypothesis ($p_H = 0.5$; Figure 2.1).

In this case, we can use the known probabilities of the binomial distribution to calculate our P-value. We want to know the probability of obtaining a result at least as extreme as our data when drawing from a binomial distribution with parameters $p = 0.5$ and $n = 100$. We calculate the area of this distribution that lies to the right of 63. This area, $P = 0.003$, can be obtained either from a table, from statistical software, or by using a relatively simple calculation. The value, 0.003, represents the probability of obtaining at least 63 heads out of 100 trials with $p_H = 0.5$. This number is the P-value from our binomial test. Because we only calculated the area of our null distribution in one tail (in this case, the right, where values are greater than or equal to 63), then this is actually a one-tailed test, and we are only considering part of our null hypothesis where $p_H > 0.5$. Such an approach might be suitable in some cases, but more typically we need to multiply this number by 2 to get a two-tailed test. By doing so, our P-value of 0.006 includes the possibility of results as extreme as our test statistic in either direction, either too many or too few heads. Since $P < 0.05$ we reject the null hypothesis, and conclude that we have an unfair lizard.

In biology, null hypotheses play a critical role in many statistical analyses. So why not end this chapter now? One issue is that biological null hypotheses are almost always uninteresting. They often describe the situation where patterns in the data occur only by chance. However, if you are comparing living species to each other, there are almost always some differences between them. In fact, for biology, null hypotheses are quite often obviously false. For example, two different species living in different habitats are not identical, and if we measure them enough we will discover this fact. From this point of view, both outcomes of a standard hypothesis test are unenlightening. One either rejects a silly hy-

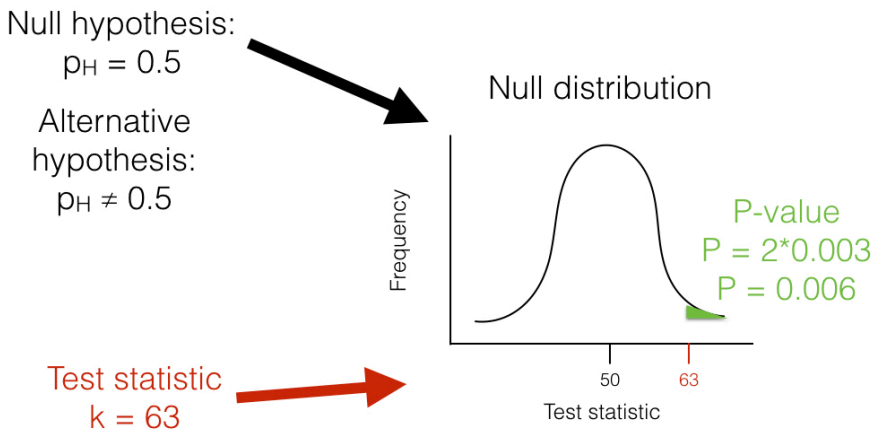

Figure 1: Figure 2.1. The unfair lizard. We use the null hypothesis to generate a null distribution for our test statistic, which in this case is a binomial distribution centered around 50. We then look at our test statistic and calculate the probability of obtaining a result at least as extreme as this value.

pothesis that was probably known to be false from the start, or one "fails to reject" this null hypothesis. There is much more information to be gained by estimating parameter values and carrying out model selection in a likelihood or Bayesian framework, as we will see below. Still, frequentist statistical approaches are common, have their place in our toolbox, and will come up in several sections of this book.

One key concept in standard hypothesis testing is the idea of statistical error. Statistical errors come in two flavors: type I and type II errors. Type I errors occur when the null hypothesis is true but the investigator mistakenly rejects it. Standard hypothesis testing controls type I errors using a parameter, α, which defines the accepted rate of type I errors. For example, if $\alpha = 0.05$, one should expect to commit a type I error about 5% of the time. When multiple standard hypothesis tests are carried out, investigators often "correct" their P-values using Bonferroni correction. If you do this, then there is only a 5% chance of a single type I error across all of the tests being considered. This singular focus on type I errors, however, has a cost. One can also commit type II errors, when the null hypothesis is false but one fails to reject it. The rate of type II errors in statistical tests can be extremely high. While statisticians do take care to create approaches that have high power, traditional hypothesis testing usually fixes type I errors at 5% while type II error rates remain unknown. There are simple ways to calculate type II error rates (e.g. power analyses) but these are only rarely carried out. Furthermore, Bonferroni correction dramatically increases the type II error rate. This is important because – as stated by Perneger (1998) – "... type II errors are no less false than type I errors."

I will cover some examples of the frequentist approach in this book, mainly when discussing traditional methods like phylogenetic independent contrasts (PICs). Also, one of the model selection approaches used frequently in this book, likelihood ratio tests, rely on a standard frequentist set-up with null and alternative hypotheses.

However, there are two good reasons to look for better ways to do comparative statistics. First, as stated above, standard methods rely on testing null hypotheses that – for evolutionary questions - are usually very likely, a priori, to be false. For a relevant example, consider a study comparing the rate of speciation between two clades of carnivores. The null hypothesis is that the two clades have exactly equal rates of speciation – which is almost certainly false, although we might question how different the two rates might be. Second, standard frequentist methods place too much emphasis on P-values and not enough on the size of statistical effects. A small P-value could reflect either a large effect or very large sample sizes or both.

In summary, frequentist statistical methods are common in comparative statistics but can be limiting. I will discuss these methods often in this book, mainly due to their prevalent use in the field. At the same time, we will look for alternatives whenever possible.

Section 2.3: Maximum likelihood

Section 2.3a: What is a likelihood?

Since all of the approaches described below involve calculating likelihoods, I will first briefly describe this concept. A good general review of likelihood is Edwards (Edwards 1992). Likelihood is defined as the probability, given a model and a set of parameter values, of obtaining a particular set of data. To calculate a likelihood, we have to consider a particular specific model that may have generated the data. That model might have parameter values that need to be specified. We can refer to this specified model as a hypothesis, H. The likelihood is then:

(eq. 2.1)

$$L(H|D) = Pr(D|H)$$

Here, L and Pr stand for likelihood and probability, D for the data, and H for the hypothesis, which again includes both the model being considered and a set of parameter values. The | symbol stands for "given," so equation 2.1 can be read as "the likelihood of the hypothesis given the data is equal to the probability of the data given the hypothesis." In other words, the likelihood represents the probability under a given model and parameter values that we would obtain the data that we actually see.

For any given model, different parameter values will generally affect the likelihood. As you might guess, we favor parameter values that give us the highest probability of obtaining the data that we see. One way to estimate parameters from data, then, is by finding the parameter values that maximize the likelihood; that is, the parameter values that give the highest likelihood, and the highest probability of obtaining the data. These estimates are then referred to as maximum likelihood (ML) estimates. In an ML framework, we suppose that the hypothesis that has the best fit to the data is the one that has the highest probability of having generated that data.

For the example above, we need to calculate the likelihood as the probability of obtaining heads 63 out of 100 lizard flips, given some model of lizard flipping. In general, we can write the likelihood for any combination of k "successes" (flips that give heads) out of n trials. We will also have one parameter, p, which will represent the probability of "success," that is, the probability that any one flip comes up heads. We can calculate the likelihood of our data using the binomial theorem:

(eq. 2.2)

$$L(H|D) = Pr(D|p) = \binom{n}{k} p_H^k (1 - p_H)^{n-k}$$

In the example given, n = 100 and k = 63, so:

(eq. 2.3)

$$L(H|D) = \binom{100}{63} p^{63}(1-p)^{37}$$

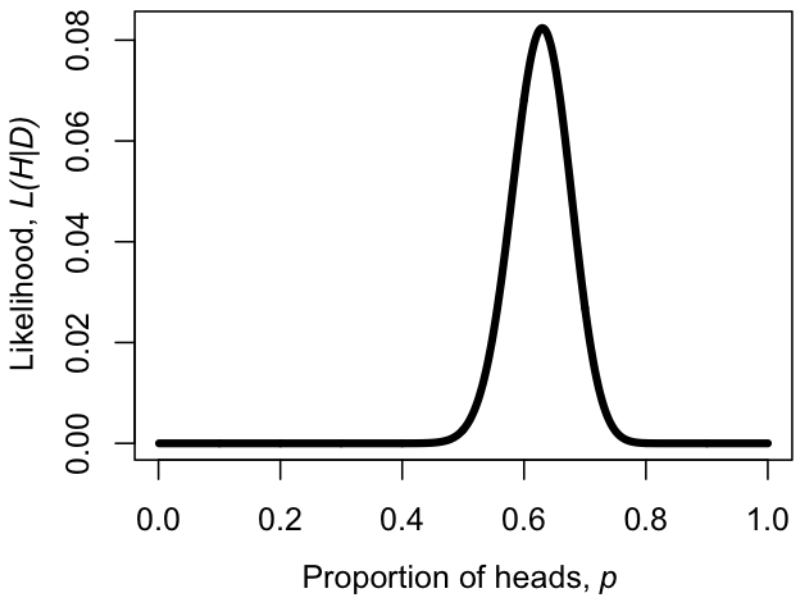

Figure 2: Figure 2.2. Likelihood surface for the parameter p_H, given a coin that has been flipped as heads 63 times out of 100.

We can make a plot of the likelihood, L, as a function of p (Figure 2.2). When we do this, we see that the maximum likelihood value of p, which we can call $\hat{p}_H$, is at $\hat{p}_H = 0.63$. This is the "brute force" approach to finding the maximum likelihood: try many different values of the parameters and pick the one with the highest likelihood. We can do this much more efficiently using numerical methods as described in later chapters in this book.

We could also have obtained the maximum likelihood estimate for p_H through differentiation. This problem is much easier if we work with the ln-likelihood rather than the likelihood itself (note that whatever value of p_H that maximizes

the likelihood will also maximize the ln-likelihood, because the log function is strictly increasing). So:

(eq. 2.4)

$$\ln L = \ln \binom{n}{k} + k \ln p_H + (n-k) \ln (1 - p_H)$$

Note that the natural log (ln) transformation changes our equation from a power function to a linear function that is easy to solve. We can differentiate:

(eq. 2.5)

$$\frac{d \ln L}{dp_H} = \frac{k}{p_H} - \frac{(n-k)}{(1-p_H)}$$

The maximum of the likelihood represents a peak, which we can find by setting the derivative $\frac{d \ln L}{dp_H}$ to zero. We then find the value of p_H that solves that equation, which will be our estimate $\hat{p}_H$. So we have:

(eq. 2.6)

$$\begin{array}{lcl} \frac{k}{\hat{p}_H} - \frac{n-k}{1-\hat{p}_H} & = & 0 \\ \frac{k}{\hat{p}_H} & = & \frac{n-k}{1-\hat{p}_H} \\ k(1-\hat{p}_H) & = & \hat{p}_H(n-k) \\ k - k\hat{p}_H & = & n\hat{p}_H - k\hat{p}_H \\ k & = & n\hat{p} \\ \hat{p}_H & = & k/n \end{array}$$

Notice that, for our simple example, $k/n = 63/100 = 0.63$, which is exactly equal to the maximum likelihood from figure 2.2.

Maximum likelihood estimates have many desirable statistical properties. It is worth noting, however, that they will not always return accurate parameter estimates, even when the data is generated under the actual model we are considering. In fact, ML parameters can sometimes be biased. To understand what this means, we need to introduce two new concepts: bias and precision. Imagine that we were to simulate datasets under some model A with parameter a. For each simulation, we then used ML to estimate the parameter $\hat{a}$ for the simulated data. The precision of our ML estimate tells us how different, on average, each of our estimated parameters $\hat{a}_i$ are from one another. Precise estimates are estimated with less uncertainty. Bias, on the other hand, measures how close our estimates $\hat{a}_i$ are to the true value a. If our ML parameter estimate is biased, then the average of the $\hat{a}_i$ will differ from the true value a. It is not uncommon for ML estimates to be biased in a way that depends on sample size, so that the estimates get closer to the truth as sample size increases, but can be

quite far off when the number of data points is small compared to the number of parameters being estimated.

In our example of lizard flipping, we estimated a parameter value of $\hat{p}_H = 0.63$. This is different from 0.5 – which was our expectation under the null hypothesis. So is this lizard fair? Or, alternatively, can we reject the null hypothesis that $p_H = 0.5$? To evaluate this, we need to use model selection.

Section 2.3b: The likelihood ratio test

Model selection involves comparing a set of potential models and using some criterion to select the one that provides the "best" explanation of the data. Different approaches define "best" in different ways. I will first discuss the simplest, but also the most limited, of these techniques, the likelihood ratio test. Likelihood ratio tests can only be used in one particular situation: to compare two models where one of the models is a special case of the other. This means that model A (the simpler model with fewer parameters) is exactly equivalent to the more complex model B with parameters restricted to certain values. For example, perhaps model B has parameters x, y, and z that can take on any values. Model A is the same as model B but with parameter z fixed at 0. That is, A is the special case of B when parameter z = 0. This is sometimes described as model A is nested within model B, since every possible version of model A is equal to a certain case of model B, but model B also includes more possibilities.

For example, consider again our example of flipping a lizard. One model is that the lizard is "fair:" that is, that the probability of heads is equal to 1/2. A different model might be that the probability of heads is some other value p, which could be 1/2, 1/3, or any other value between 0 and 1. Here, the latter (complex) model has one additional parameter, p_H, compared to the former (simple) model; the simple model is a special case of the complex model when $p_H = 1/2$.

For such nested models, one can calculate the likelihood ratio test statistic as

(eq. 2.7)

$$\Delta = 2 \cdot \ln \frac{L_1}{L_2} = 2 \cdot (\ln L_1 - \ln L_2)$$

Here, Δ is the likelihood ratio test statistic, L_2 the likelihood of the more complex (parameter rich) model, and L_1 the likelihood of the simpler model. Since the models are nested, the likelihood of the complex model will always be greater than or equal to the likelihood of the simple model; this means that the test statistic Δ will never be negative. In fact, if you ever obtain a negative likelihood ratio test statistic, something has gone wrong – either your calculations are wrong, or you have not actually found ML solutions, or the models are not actually nested.

To carry out a statistical test comparing the two models, we compare the test statistic Δ to its expectation under the null hypothesis. For likelihood ratio tests, the null hypothesis is always the simpler of the two models. When sample sizes are large, the null distribution of the likelihood ratio test statistic follows a chi-squared (χ^2) distribution with degrees of freedom equal to the difference in the number of parameters between the two models. This means that if the simpler hypothesis were true, and one carried out this test many times on large independent datasets, the test statistic would approximately follow this χ^2 distribution. To reject the simpler model, then, one compares the test statistic with a critical value derived from the appropriate χ^2 distribution. If the test statistic is larger than the critical value, one rejects the null hypothesis. Otherwise, we fail to reject the null hypothesis. In this case, we only need to consider one tail of the χ^2 test, as every deviation from the null model will push us towards higher Δ values and towards the right tail of the distribution.

For the lizard flip example above, we can calculate the ln-likelihood under a hypothesis of $p_H = 0.5$ as:

(eq. 2.8)

$$\begin{aligned} \ln L_1 &= \ln\left(\tfrac{100}{63}\right) + 63 \cdot \ln 0.5 + (100 - 63) \cdot \ln(1 - 0.5) \\ \ln L_1 &= -5.92 \end{aligned}$$

We can compare this to the likelihood of our maximum-likelihood estimate :

(eq. 2.9)

$$\begin{aligned} \ln L_2 &= \ln\left(\tfrac{100}{63}\right) + 63 \cdot \ln 0.63 + (100 - 63) \cdot \ln(1 - 0.63) \\ \ln L_2 &= -2.50 \end{aligned}$$

We then calculate the likelihood ratio test statistic:

(eq. 2.10)

$$\begin{aligned} \Delta &= 2 \cdot (\ln L_2 - \ln L_1) \\ \Delta &= 2 \cdot (-2.50 - -5.92) \\ \Delta &= 6.84 \end{aligned}$$

If we compare this to a χ^2 distribution with one d.f., we find that $P = 0.009$. Because this P-value is less than the threshold of 0.05, we reject the null hypothesis, and support the alternative. We conclude that this is not a fair lizard.

Although described above in terms of two competing hypotheses, likelihood ratio tests can be applied to more complex situations with more than two competing models. For example, if all of the models form a sequence of increasing complexity, with each model a special case of the next more complex model, one can compare each pair of hypotheses in sequence, stopping the first time the

test statistic is non-significant. Alternatively, in some cases, hypotheses can be placed in a bifurcating choice tree, and one can proceed from simple to complex models down a particular path of paired comparisons of nested models. This approach is commonly used to select models of DNA sequence evolution.

Section 2.3c: The Akaike information criterion (AIC)

You might have noticed that the likelihood ratio test described above has some limitations. Especially for models involving more than one parameter, approaches based on likelihood ratio tests can only do so much. For example, one can compare a series of models, some of which are nested within others, using an ordered series of likelihood ratio tests. However, results will often depend strongly on the order in which tests are carried out. Furthermore, often we want to compare models that are not nested, as required by likelihood ratio tests. For these reasons, another approach, based on the Akaike Information Criterion (AIC), can be useful.

The AIC value for a particular model is a simple function of the likelihood L and the number of parameters k:

(eq. 2.11)

\[AIC = 2k - 2 \ln\{L\} \]

This function that balances the likelihood of the model and the number of parameters estimated in the process of fitting the model to the data. One can think of the AIC criterion as identifying the model that provides the most efficient way to describe patterns in the data with few parameters. However, this shorthand description of AIC does not capture the actual mathematical and philosophical justification for equation (2.11). In fact, this equation is not arbitrary; instead, it comes from information theory (for more information, see Burnham and Anderson 2003).

The AIC equation (2.11) above is only valid for quite large sample sizes relative to the number of parameters being estimated (for n samples and k parameters, $n/k > 40$). Most empirical data sets include fewer than 40 independent data points per parameter, so a small sample size correction should be employed:

(eq. 2.12)

$$AIC_C = AIC + \frac{2k(k+1)}{n-k-1}$$

This correction penalizes models that have small sample sizes relative to the number of values that are too close; that is, models where there are nearly as many parameters as data points. As noted by Burnham and Anderson (2003), this correction has little effect if sample sizes are large, and so provides a robust

way to correct for possible bias in data sets of any size. I recommend always using the small sample size correction when calculating AIC values.

To select among models, one can then compare their AIC_c values, and choose the model with the smallest value. It is easier to make comparisons in AIC_c scores between models by calculating the difference, ΔAIC_c. For example, if you are comparing a set of models, you can calculate ΔAIC_c for model i as:

(eq. 2.13)

$$\Delta AIC_{c_i} = AIC_{c_i} - AIC_{c_{min}}$$

where AIC_{c_i} is the AIC_c score for model i and $AIC_{c_{min}}$ is the minimum AIC_c score across all of the models.

As a broad rule of thumb for comparing AIC values, any model with a ΔAIC_{c_i} of less than four is roughly equivalent to the model with the lowest AIC_c value. Models with ΔAIC_{c_i} between 4 and 8 have little support in the data, while any model with a ΔAIC_{c_i} greater than 10 can safely be ignored.

Additionally, one can calculate the relative likelihood for each model using Akaike weights. The weight for model i compared to a set of competing models is calculated as:

(eq. 2.14)

$$w_i = \frac{e^{-\Delta AIC_{c_i}/2}}{\sum_i e^{-\Delta AIC_{c_i}/2}}$$

The weights for all models under consideration sum to 1, so the w_i for each model can be viewed as an estimate of the level of support for that model in the data compared to the other models being considered.

Returning to our example of lizard flipping, we can calculate AIC_c scores for our two models as follows:

(eq. 2.15)

$$\begin{aligned} AIC_1 &= 2k_1 - 2lnL_1 = 2 \cdot 0 - 2 \cdot -5.92 \\ AIC_1 &= 11.8 \\ AIC_2 &= 2k_2 - 2lnL_2 = 2 \cdot 1 - 2 \cdot -2.50 \\ AIC_2 &= 7.0 \end{aligned}$$

Our example is a bit unusual in that model one has no estimated parameters; this happens sometimes but is not typical for biological applications. We can

correct these values for our sample size, which in this case is $n = 100$ lizard flips:

(eq. 2.16)

$$\begin{aligned} AIC_{c_1} &= AIC_1 + \tfrac{2k_1(k_1+1)}{n-k_1-1} \\ AIC_{c_1} &= 11.8 + \tfrac{2 \cdot 0(0+1)}{100-0-1} \\ AIC_{c_1} &= 11.8 \\ AIC_{c_2} &= AIC_2 + \tfrac{2k_2(k_2+1)}{n-k_2-1} \\ AIC_{c_2} &= 7.0 + \tfrac{2 \cdot 1(1+1)}{100-1-1} \\ AIC_{c_2} &= 7.0 \end{aligned}$$

Notice that, in this particular case, the correction did not affect our AIC values, at least to one decimal place. This is because the sample size is large relative to the number of parameters. Note that model 2 has the smallest AIC_c score and is thus the model that is best supported by the data. Noting this, we can now convert these AIC_c scores to a relative scale:

(eq. 2.17)

$$\begin{aligned} \Delta AIC_{c_1} &= AIC_{c_1} - AICc_{min} \\ &= 11.8 - 7.0 \\ &= 4.8 \end{aligned}$$

$$\begin{aligned} \Delta AIC_{c_2} &= AIC_{c_2} - AICc_{min} \\ &= 7.0 - 7.0 \\ &= 0 \end{aligned}$$

Note that the ΔAIC_{c_i} for model 1 is greater than four, suggesting that this model (the "fair" lizard) has little support in the data. Finally, we can use the relative AICc scores to calculate Akaike weights:

(eq. 2.18)

$$\begin{aligned} \textstyle\sum_i e^{-\Delta_i/2} &= e^{-\Delta_1/2} + e^{-\Delta_2/2} \\ &= e^{-4.8/2} + e^{-0/2} \\ &= 1.09 \end{aligned}$$

$$
\begin{aligned}
w_1 &= \frac{e^{-\Delta AICc_1/2}}{\sum_i e^{-\Delta AICc_i/2}} \\
&= \frac{0.09}{1.09} \\
&= 0.08
\end{aligned}
$$

$$
\begin{aligned}
w_2 &= \frac{e^{-\Delta AICc_2/2}}{\sum_i e^{-\Delta AICc_i/2}} \\
&= \frac{1.00}{1.09} \\
&= 0.92
\end{aligned}
$$

Our results are again consistent with the results of the likelihood ratio test. The relative likelihood of an unfair lizard is 0.92, and we can be quite confident that our lizard is not a fair flipper.

AIC weights are also useful for another purpose: we can use them to get model-averaged parameter estimates. These are parameter estimates that are combined across different models proportional to the support for those models. As a thought example, imagine that we are considering two models, A and B, for a particular dataset. Both model A and model B have the same parameter p, and this is the parameter we are particularly interested in. In other words, we do not know which model is the best model for our data, but what we really need is a good estimate of p. We can do that using model averaging. If model A has a high AIC weight, then the model-averaged parameter estimate for p will be very close to our estimate of p under model A; however, if both models have about equal support then the parameter estimate will be close to the average of the two different estimates. Model averaging can be very useful in cases where there is a lot of uncertainty in model choice for models that share parameters of interest. Sometimes the models themselves are not of interest, but need to be considered as possibilities; in this case, model averaging lets us estimate parameters in a way that is not as strongly dependent on our choice of models.

Section 2.4: Bayesian statistics

Section 2.4a: Bayes Theorem

Recent years have seen tremendous growth of Bayesian approaches in reconstructing phylogenetic trees and estimating their branch lengths. Although there are currently only a few Bayesian comparative methods, their number will certainly grow as comparative biologists try to solve more complex problems. In a Bayesian framework, the quantity of interest is the posterior probability, calculated using Bayes' theorem:

(eq. 2.19)

$$Pr(H|D) = \frac{Pr(D|H) \cdot Pr(H)}{Pr(D)}$$

The benefit of Bayesian approaches is that they allow us to estimate the probability that the hypothesis is true given the observed data, $Pr(H|D)$. This is really the sort of probability that most people have in mind when they are thinking about the goals of their study. However, Bayes theorem also reveals a cost of this approach. Along with the likelihood, $Pr(D|H)$, one must also incorporate prior knowledge about the probability that any given hypothesis is true - $Pr(H)$. In Bayesian statistics one must quantify the prior belief that a hypothesis is true, even before consideration of the data at hand. In practice, scientists often seek to use "uninformative" priors that have little influence on the posterior distribution - although even the term "uninformative" can be confusing, because the prior is an integral part of a Bayesian analysis. The term $Pr(D)$ is also an important part of Bayes theorem, and can be calculated as the probability of obtaining the data integrated over the prior distributions of the parameters:

(eq. 2.20)

$$Pr(D) = \int_H Pr(H|D)Pr(H)dH$$

However, $Pr(D)$ is constant when comparing the fit of different models for a given data set and thus has no influence on Bayesian model selection under most circumstances (and all the examples in this book).

In our example of lizard flipping, we can do an analysis in a Bayesian framework. For model 1, there are no free parameters. Because of this, $Pr(H) = 1$ and $Pr(D|H) = P(D)$, so that $Pr(H|D) = 1$. This may seem strange but what the result means is that our data has no influence on the structure of the model. We do not learn anything about a model with no free parameters by collecting data!

If we consider model 2 above, the parameter p_H must be estimated. We can set a uniform prior between 0 and 1 for p_H, so that $f(p_H) = 1$ for all p_H in the interval [0,1]. We can also write this as "our prior for p_h is U(0,1)". Then:

(eq. 2.21)

$$Pr(H|D) = \frac{Pr(D|H) \cdot Pr(H)}{Pr(D)} = \frac{P(k|p_H, N)f(p_H)}{\int_0^1 P(k|p_H, N)f(p_h)dp_H}$$

Next we note that $Pr(D|H)$ is the likelihood of our data given the model, which is already stated above as equation 2.2. Plugging this into our equation, we have:

(eq. 2.22)

$$Pr(H|D) = \frac{\binom{N}{k} p_H^k (1-p_H)^{N-k}}{\int_0^1 \binom{N}{k} p_H^k (1-p_H)^{N-k} dp_H}$$

This ugly equation is actually a beta distribution, which can be expressed more simply as:

(eq. 2.23)

$$Pr(H|D) = \frac{(N+1)!}{k!(N-k)!} p_H^k (1-p_H)^{N-k}$$

We can compare this posterior distribution of our parameter estimate, p_H, given the data, to our uniform prior (Figure 2.3). If you inspect this plot, you see that the posterior distribution is very different from the prior – that is, the data have changed our view of the values that parameters should take.

As you can see from this example, Bayes theorem lets us combine our prior belief about parameter values with the information from the data in order to obtain a posterior. These posterior distributions are very easy to interpret, as they express the probability of the model parameters given our data. However, that clarity comes at a cost of requiring an explicit prior. Later in the book we will learn how to use this feature of Bayesian statistics to our advantage when we actually do have some prior knowledge about parameter values.

Section 2.4b: Bayesian MCMC

The other main tool in the toolbox of Bayesian comparative methods is the use of Markov-chain Monte Carlo (MCMC) tools to calculate posterior probabilities. MCMC techniques use an algorithm that uses a "chain" of calculations to sample the posterior distribution. MCMC requires calculation of likelihoods but not complicated mathematics (e.g. integration of probability distributions), and so represents a more flexible approach to Bayesian computation. Frequently, the integrals in equation 2.21 are intractable, so that the most efficient way to fit Bayesian models is by using MCMC. Also, setting up an MCMC is, in my experience, easier than people expect!

An MCMC analysis requires that one constructs and samples from a Markov chain. A Markov chain is a random process that changes from one state to another with certain probabilities that depend only on the current state of the system, and not what has come before. A simple example of a Markov chain is the movement of a playing piece in the game Chutes and Ladders; the position of the piece moves from one square to another following probabilities given by

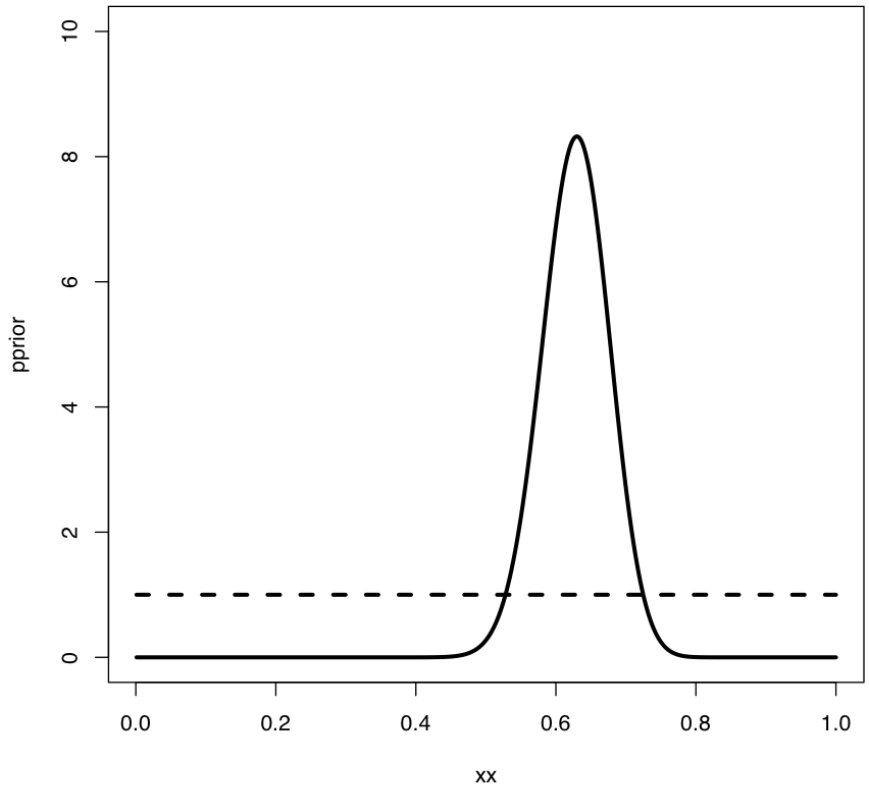

Figure 3: Figure 2.3. Bayesian prior (dotted line) and posterior (solid line) distributions for lizard flipping.

the dice and the layout of the game board. The movement of the piece from any square on the board does not depend on how the piece got to that square.

Some Markov chains have an equilibrium distribution, which is a stable probability distribution of the model's states after the chain has run for a very long time. For Bayesian analysis, we use a technique called a Metropolis-Hasting algorithm to construct a special Markov chain that has an equilibrium distribution that is the same as the Bayesian posterior distribution of our statistical model. Then, using a random simulation on this chain (this is the Markov-chain Monte Carlo, MCMC), we can sample from the posterior distribution of our model.

The following algorithm uses a Metropolis-Hastings algorithm to carry out a Bayesian MCMC analysis with one free parameter:

---

1. Get a starting parameter value.

Sample a starting parameter value, p, from the prior distribution.

2. Propose a new parameter.

Given the current parameter value, p, select a new proposed parameter value, p, using the proposal density $Q(p'|p)$.

3. Calculate three ratios.

a. The prior odds ratio.

This is the ratio of the probability of drawing the parameter values p and p from the prior.

(eq. 2.24)

$$R_{prior} = \frac{P(p')}{P(p)}$$

b. The proposal density ratio.

This is the ratio of probability of proposals going from p to p and the reverse. Often, one can construct a proposal density that is symmetrical, so that $Q(p'|p) = Q(p|p')$ and $a_2 = 1$.

(eq. 2.25)

$$R_{proposal} = \frac{Q(p'|p)}{Q(p|p')}$$

c. The likelihood ratio.

This is the ratio of probabilities of the data given the two different parameter values.

(eq. 2.26)

$$R_{likelihood} = \frac{L(p'|D)}{L(p|D)} = \frac{P(D|p')}{P(D|p)}$$

4. Multiply.

Find the product of the prior odds, proposal density ratio, and the likelihood ratio:

(eq. 2.27)

$$R_{accept} = R_{prior} \cdot R_{proposal} \cdot R_{likelihood}$$

5. Accept or reject.

Draw a random number x from a uniform distribution between 0 and 1. If $x < R_{accept}$, accept the proposed value of p; otherwise reject, and retain the current value p.

6. Repeat.

Repeat steps 2-5 a large number of times.

Carrying out these steps, one obtains a set of parameter values, p_i, where i is from 1 to the total number of generations in the MCMC. Typically, the chain has a "burn-in" period at the beginning. This is the time before the chain has reached a stationary distribution, and can be observed when parameter values show trends through time and the likelihood for models has yet to plateau. If you eliminate this "burn-in" period, then you can treat the rest of the chain as a sample from the posterior distribution, and summarize it in a variety of ways; for example, by calculating a mean, 95% confidence interval, or plotting a histogram.

We can apply this algorithm to our coin-flipping example. We will consider the same prior distribution, $U(0,1)$, for the parameter p. We will also define a proposal density, $Q(p'|p)$ $U(p-\epsilon, p+\epsilon)$. That is, we will add or subtract a small number ($\epsilon \leq 0.01$) to generate proposed values of p given p.

To start the algorithm, we draw a value of p from the prior. Let's say for illustrative purposes that the value we draw is 0.60. This becomes our current parameter estimate. For step two, we propose a new value, p, by drawing from our proposal distribution. We can use $\epsilon = 0.01$ so the proposal distribution becomes $U(0.59, 0.61)$. Let's suppose that our new proposed value $p = 0.595$.

We then calculate our three ratios. Here things are simpler than you might have expected for two reasons. First, recall that our prior probability distribution is $U(0,1)$. The density of this distribution is a constant (1.0) for all values of p and p. Because of this, the prior odds ratio is always:

(eq. 2.28)

$$R_{prior} = \frac{P(p')}{P(p)} = \frac{1}{1} = 1$$

Similarly, because our proposal distribution is symmetrical, $Q(p'|p) = Q(p|p')$ and $R_{proposal} = 1$. That means that we only need to calculate the likelihood ratio, $R_{likelihood}$ for p and p. We can do this by plugging our values for p (or p) into equation 2.2:

(eq. 2.29)

$$P(D|p) = \binom{N}{k} p^k (1-p)^{N-k} = \binom{100}{63} 0.6^6 3(1-0.6)^{100-63} = 0.068$$

Likewise, (eq. 2.30)

$$P(D|p') = \binom{N}{k} p'^k (1-p')^{N-k} = \binom{100}{63} 0.595^6 3(1-0.595)^{100-63} = 0.064$$

The likelihood ratio is then:

(eq. 2.31)

$$R_{likelihood} = \frac{P(D|p')}{P(D|p)} = \frac{0.064}{0.068} = 0.94$$

We can now calculate $R_{accept} = R_{prior} \cdot R_{proposal} \cdot R_{likelihood} = 1 \cdot 1 \cdot 0.94 = 0.94$. We next choose a random number between 0 and 1 – say that we draw $x = 0.34$.

We then notice that our random number x is less than or equal to R_{accept}, so we accept the proposed value of p. If the random number that we drew were greater than 0.94, we would reject the proposed value, and keep our original parameter value $p = 0.60$ going into the next generation.

If we repeat this procedure a large number of times, we will obtain a long chain of values of p. You can see the results of such a run in Figure 2.4. In panel A, I have plotted the likelihoods for each successive value of p. You can see that the likelihoods increase for the first ~1000 or so generations, then reach a plateau around $lnL = -3$. Panel B shows a plot of the values of p, which rapidly converge to a stable distribution around $p = 0.63$. We can also plot a histogram of these posterior estimates of p. In panel C, I have done that – but with a twist. Because the MCMC algorithm creates a series of parameter estimates, these numbers show autocorrelation – that is, each estimate is similar to estimates that come just before and just after. This autocorrelation can cause problems for data analysis. The simplest solution is to subsample these values, picking only, say, one value every 100 generations. That is what I have done in the histogram in panel C. This panel also includes the analytic posterior distribution that we calculated above – notice how well our Metropolis-Hastings algorithm did in reconstructing this distribution!

This simple example glosses over some of the details of MCMC algorithms, but we will get into those details later, and there are many other books that treat this topic in great depth (e.g. Christensen et al. 2010). The point is that we can solve some of the challenges involved in Bayesian statistics using numerical "tricks" like MCMC, that exploit the power of modern computers to fit models and estimate model parameters.

Section 2.4c: Bayes factors

Now that we know how to use data and a prior to calculate a posterior distribution, we can move to the topic of model selection. We already learned one general method for model selection using AIC. We can also do model selection in a Bayesian framework. The simplest way is to calculate and then compare the posterior probabilities for a set of models under consideration. One can do this by calculating Bayes factors:

(eq. 2.32)

$$B_{12} = \frac{Pr(D|H_1)}{Pr(D|H_2)}$$

Bayes factors are ratios of the marginal likelihoods $P(D|H)$ of two competing models. They represent the probability of the data averaged over the posterior distribution of parameter estimates. It is important to note that these marginal

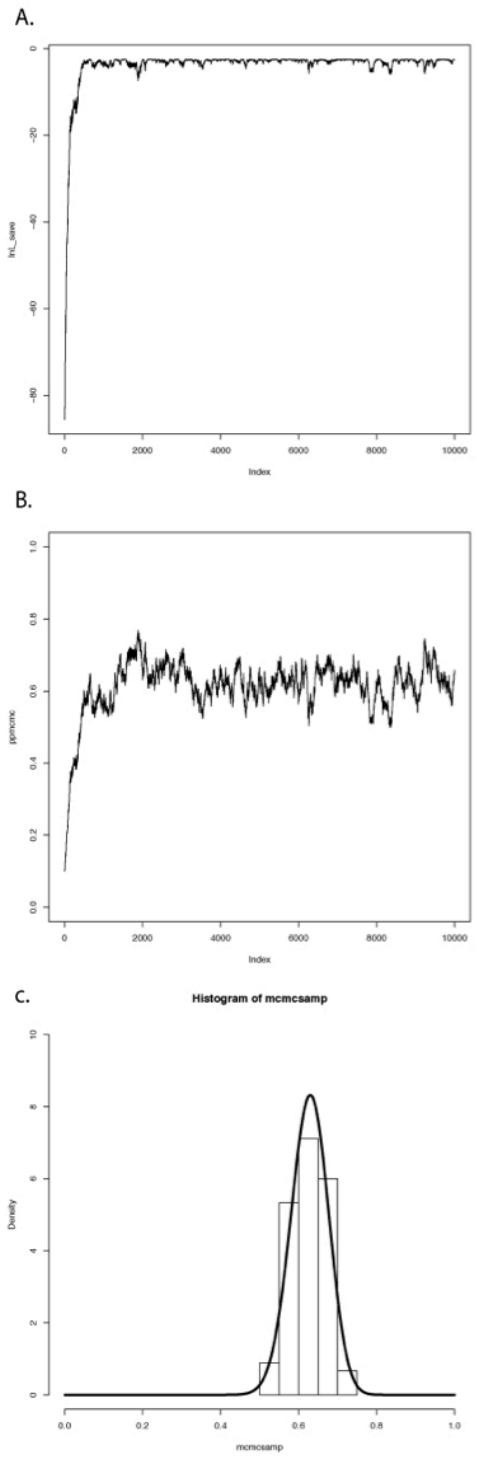

Figure 4: Figure 2.4. Bayesian MCMC from lizard flipping example.

likelihoods are different from the likelihoods used above for AIC model comparison in an important way. With AIC and other related tests, we calculate the likelihoods for a given model and a particular set of parameter values – in the coin flipping example, the likelihood for model 2 when $p_H = 0.63$. By contrast, Bayes factors' marginal likelihoods give the probability of the data averaged over all possible parameter values for a model, weighted by their prior probability.

Because of the use of marginal likelihoods, Bayes factor allows us to do model selection in a way that accounts for uncertainty in our parameter estimates – again, though, at the cost of requiring explicit prior probabilities for all model parameters. Such comparisons can be quite different from likelihood ratio tests or comparisons of AIC_c scores. Bayes factors represent model comparisons that integrate over all possible parameter values rather than comparing the fit of models only at the parameter values that best fit the data. In other words, AIC_c scores compare the fit of two models given particular estimated values for all of the parameters in each of the models. By contrast, Bayes factors make a comparison between two models that accounts for uncertainty in their parameter estimates. This will make the biggest difference when some parameters of one or both models have relatively wide uncertainty. If all parameters can be estimated with precision, results from both approaches should be similar.

Calculation of Bayes factors can be quite complicated, requiring integration across probability distributions. In the case of our coin-flipping problem, we have already done that to obtain the beta distribution in equation 2.22. We can then calculate Bayes factors to compare the fit of two competing models. Let's compare the two models for coin flipping considered above: model 1, where $p_H = 0.5$, and model 2, where $p_H = 0.63$. Then:

(eq. 2.33)

$$\begin{aligned}
Pr(D|H_1) &= \tbinom{100}{63} 0.5^0 .63(1-0.5)^{100-63} \\
&= 0.00270 \\
Pr(D|H_2) &= \int_{p=0}^{1} \tbinom{100}{63} p^6 3(1-p)^{100-63} \\
&= \tbinom{100}{63} \beta(38, 64) \\
&= 0.0099 \\
B_{12} &= \tfrac{0.0099}{0.00270} \\
&= 3.67
\end{aligned}$$

In the above example, $\beta(x, y)$ is the Beta function. Our calculations show that the Bayes factor is 3.67 in favor of model 2 compared to model 1. This is typically interpreted as substantial (but not decisive) evidence in favor of model 2. Again, we can be reasonably confident that our lizard is not a fair flipper.

In the lizard flipping example we can calculate Bayes factors exactly because we know the solution to the integral in equation 2.33. However, if we don't

know how to solve this equation (a typical situation in comparative methods), we can still approximate Bayes factors from our MCMC runs. Methods to do this, including arrogance sampling and stepping stone models, are complex and beyond the scope of this book. However, one common method for approximating Bayes Factors involves calculating the harmonic mean of the likelihoods over the MCMC chain for each model. The ratio of these two likelihoods is then used as an approximation of the Bayes factor (Newton and Raftery 1994). Unfortunately, this method is extremely unreliable, and probably should never be used (see this blog post for more details).

Section 2.5: AIC versus Bayes

Before I conclude this section, I want to highlight another difference in the way that *AIC* and Bayes approaches deal with model complexity. This relates to a subtle philosophical distinction that is controversial among statisticians themselves so I will only sketch out the main point; see a real statistics book like Burnham and Anderson (2003) or Gelman et al. (2013) for further details. When you compare Bayes factors, you assume that one of the models you are considering is actually the true model that generated your data, and calculate posterior probabilities based on that assumption. By contrast, *AIC* assumes that reality is more complex than any of your models, and you are trying to identify the model that most efficiently captures the information in your data. That is, even though both techniques are carrying out model selection, the basic philosophy of how these models are being considered is very different: choosing the best of several simplified models of reality, or choosing the correct model from a set of alternatives.

The debate between Bayesian and likelihood-based approaches often centers around the use of priors in Bayesian statistics, but the distinction between models and "reality" is also important. More specifically, it is hard to imagine a case in comparative biology where one would be justified in the Bayesian assumption that one has identified the true model that generated the data. This also explains why *AIC*-based approaches typically select more complex models than Bayesian approaches. In an *AIC* framework, one assumes that reality is very complex and that models are approximations; the goal is to figure out how much added model complexity is required to efficiently explain the data. In cases where the data are actually generated under a very simple model, *AIC* may err in favor of overly complex models. By contrast, Bayesian analyses assume that one of the models being considered is correct. This type of analysis will typically behave appropriately when the data are generated under a simple model, but may be unpredictable when data are generated by processes that are not considered by any of the models. However, Bayesian methods account for uncertainty much better than AIC methods, and uncertainty is a fundamental aspect of phylogenetic comparative methods.

In summary, Bayesian approaches are useful tools for comparative biology, es-

pecially when combined with MCMC computational techniques. They require specification of a prior distribution and assume that the "true" model is among those being considered, both of which can be drawbacks in some situations. A Bayesian framework also allows us to much more easily account for phylogenetic uncertainty in comparative analysis. Many comparative biologists are pragmatic, and use whatever methods are available to analyze their data. This is a reasonable approach but one should remember the assumptions that underlie any statistical result.

Section 2.6: Models and comparative methods

For the rest of this book I will introduce several models that can be applied to evolutionary data. I will discuss how to simulate evolutionary processes under these models, how to compare data to these models, and how to use model selection to discriminate amongst them. In each section, I will describe standard statistical tests (when available) along with ML and Bayesian approaches.

One theme in the book is that I emphasize fitting models to data and estimating parameters. I think that this approach is very useful for the future of the field of comparative statistics for three main reasons. First, it is flexible; one can easily compare a wide range of competing models to your data. Second, it is extendable; one can create new models and automatically fit them into a preexisting framework for data analysis. Finally, it is powerful; a model fitting approach allows us to construct comparative tests that relate directly to particular biological hypotheses.

Chapter 2 References

Burnham, K. P., and D. R. Anderson. 2003. Model selection and multimodel inference: A practical Information-Theoretic approach. Springer Science & Business Media.

Edwards, A. W. F. 1992. Likelihood. Johns Hopkins University Press, Baltimore.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. 2013. Bayesian data analysis, third edition. Chapman; Hall/CRC.

Newton, M. A., and A. E. Raftery. 1994. Approximate bayesian inference with the weighted likelihood bootstrap. J. R. Stat. Soc. Series B Stat. Methodol. 56:3–48.

Perneger, T. V. 1998. What's wrong with bonferroni adjustments. BMJ 316:1236–1238.

pdf version

Chapter 3: Introduction to Brownian Motion

Section 3.1: Introduction

Squamates, the group that includes snakes and lizards, is exceptionally diverse. This clade, which is between 150 and 210 million years old (Hedges and Kumar 2009), includes species that are very large and very small; herbivores and carnivores; species with legs and species that are legless. How did that diversity of species' traits come to be? How did these characters first come to be, and how often did they change to explain the diversity that we see on earth today? In this chapter, we will begin to discuss models for the evolution of species' traits.

Imagine that you want to use statistical approaches to understand how traits change through time. To do that, you need to have an exact mathematical specification of how evolution takes place. Obviously there are a wide variety of models of trait evolution, from simple to complex. For example, you might create a model where a trait starts with a certain value and has some constant probability of changing in any unit of time. Alternatively, you might make a model that is more explicit, and considers a large set of individuals in a population. You could assign genotypes to each individual and allow the population to change through reproduction and natural selection. In this chapter – and in comparative methods as a whole – the models we will consider will be much closer to the first of these two models. However, there are still important connections between these simple models and more realistic models of trait evolution. (see chapter 5).

In the next six chapters, I will discuss models for two different types of characters. In chapters three, four, and five, I will consider traits that follow continuous distributions – that is, traits that can have real-numbered values. For example, body mass in kilograms is a continuous character. I will discuss the most commonly used model for these continuous characters, Brownian motion, in this chapter and the next, and go beyond Brownian motion in chapter five. In chapters six, seven, and eight, I will cover discrete characters, characters that can occupy one of a number of distinct character states (for example, species of squamates can either be legless or have legs).

Section 3.2: Properties of Brownian Motion

We can use Brownian motion to model the evolution of a continuously valued trait through time. Brownian motion is an example of a "random walk" model because the trait value changes randomly, in both direction and distance, over any time interval.

The statistical process of Brownian motion was originally invented to describe the motion of particles suspended in a fluid. To me this is a bit hard to picture, but the logic applies equally well to the movement of a large ball over a crowd in a stadium. When the ball is over the crowd, people push on it from many directions. The sum of these many small forces determine the movement of the ball. Again, the movement of the ball – considered in two dimensions to describe movement both across and up and down the stadium rows – can be modeled using Brownian motion.

The core idea of this example is that the motion of the object is due to the sum of a large number of very small, random forces. This idea is a key part of biological models of evolution under Brownian motion. It is worth mentioning that even though Brownian motion involves change that has a strong random component, it is incorrect to equate Brownian motion models with models of pure genetic drift (as explained in more detail below).

Brownian motion is a popular model in comparative biology because it captures the way traits might evolve under a reasonably wide range of scenarios. However, perhaps the main reason for the dominance of Brownian motion as a model is that it has some very convenient statistical properties that allow relatively simple analyses and calculations on trees. I will use some simple simulations to show how the Brownian motion model behaves. I will then list the three critical statistical properties of Brownian motion, and explain how we can use these properties to apply Brownian motion models to phylogenetic comparative trees.

When we model evolution using Brownian motion, we are typically discussing the dynamics of the mean character value, which we will denote as $\bar{z}$, in a population. That is, we imagine that you can measure a sample of the individuals in a population and estimate the mean average trait value. We will denote the mean trait value at some time t as $\bar{z}(t)$. We can then model the mean trait value through time with a Brownian motion process.

Brownian motion models can be completely described by two parameters. The first is the starting value of the population mean trait, $\bar{z}(0)$. This is the mean trait value that is seen in the ancestral population at the start of the simulation, before any trait change occurs. The second parameter of Brownian motion is the evolutionary rate parameter, σ^2. This parameter determines how fast traits will randomly walk through time.

At the core of Brownian motion is the normal distribution. You might know that a normal distribution can be described by two parameters, the mean and variance. We can simulate change under Brownian motion model by drawing from normal distributions. In particular, changes in trait values over any interval of time are always drawn from a normal distribution with mean 0 and variance proportional to the product of the rate of evolution and the length of time (variance = $\sigma^2 t$). Another way to say this is that the expected change under a Brownian motion model follows a normal distribution with mean 0 and variance proportional to the elapsed time.

A few plots will illustrate the behavior of Brownian motion. Figure 3.1 shows sets of Brownian motion run over three different time periods (t = 100, 500, and 1000) with the same starting value $\bar{z}(0) = 0$ and rate parameter $\sigma^2 = 1$. Each panel of the figure shows 100 simulations of the process over that time period. You can see that the tip values look like normal distributions. Furthermore, the variance among separate runs of the process increases linearly with time. This among-run variance is greatest over the longest time intervals. It is this variance, the variation among many independent runs of the same evolutionary process, that we will consider throughout the next section.

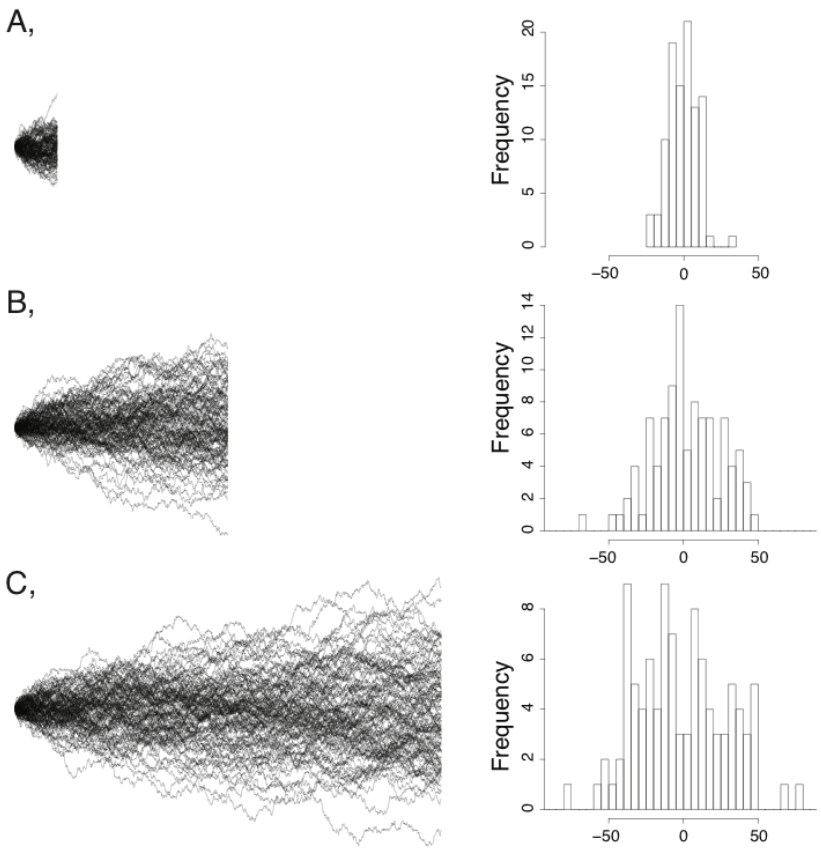

Figure 1:

Figure 3.1. Examples of Brownian motion. Each plot shows 100 replicates of simulated Brownian motion with a common starting value and the same rate parameter $\sigma^2 = 1$. Simulations were run for three different times: (A) 10, (B) 50, and (C) 100 time units. The right-hand column shows a histogram of the

distribution of ending values for each set of 100 simulations.

Imagine that we run a Brownian motion process over a given time interval many times, and save the trait values at the end of each of these simulations. We can then create a statistical distribution of these character states. It might not be obvious from figure 3.1, but the distributions of possible character states at any time point in a Brownian walk is normal. This is illustrated in figure 3.2, which shows the distribution of traits from 100,000 simulations with $\sigma^2 = 1$ and $t = 100$. The tip characters from all of these simulations follow a normal distribution with mean equal to the starting value, $\bar{z}(0) = 0$, and a variance of $\sigma^2 t = 100$.

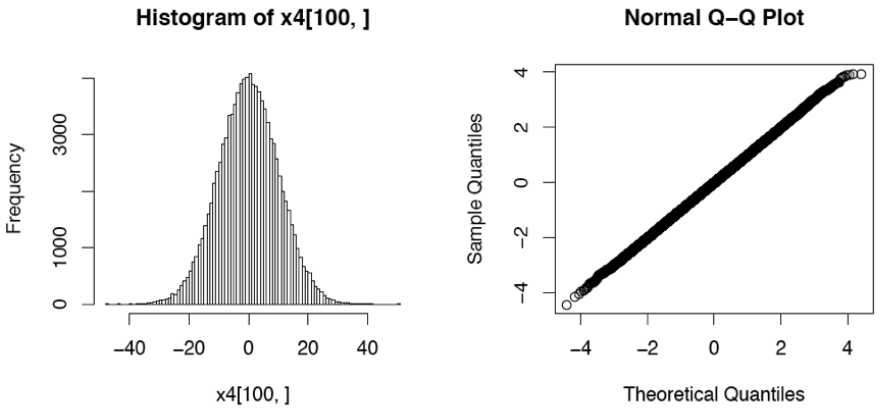

Figure 2:

Figure 3.2. Ending character values from of 100,000 Brownian motion simulations with $\Theta = 0$, $t = 100$, and $\sigma^2 = 1$. Panel (A) shows a histogram of the outcome of these simulations, while panel (B) shows a normal Q-Q plot for these data. If the data follow a normal distribution, the points in the Q-Q plot should form a straight line.

Figure 3.3 shows how rate parameter σ^2 affects the rate of spread of Brownian walks. The panels show sets of 100 Brownian motion simulations run over 1000 time units for $\sigma^2 = 1$ (Panel A), $\sigma^2 = 5$ (Panel B), and $\sigma^2 = 25$ (Panel C). You can see that simulations with a higher rate parameter create a larger spread of trait values per unit time.

Figure 3.3. Examples of Brownian motion. Each plot shows 100 replicates of simulated Brownian motion with a common starting value and the same time interval $t = 100$. The rate parameter σ^2 varies across the panels: (A) $\sigma^2 = 1$ (B) $\sigma^2 = 10$, and (C) $\sigma^2 = 25$. The right-hand column shows a histogram of the distribution of ending values for each set of 100 simulations.

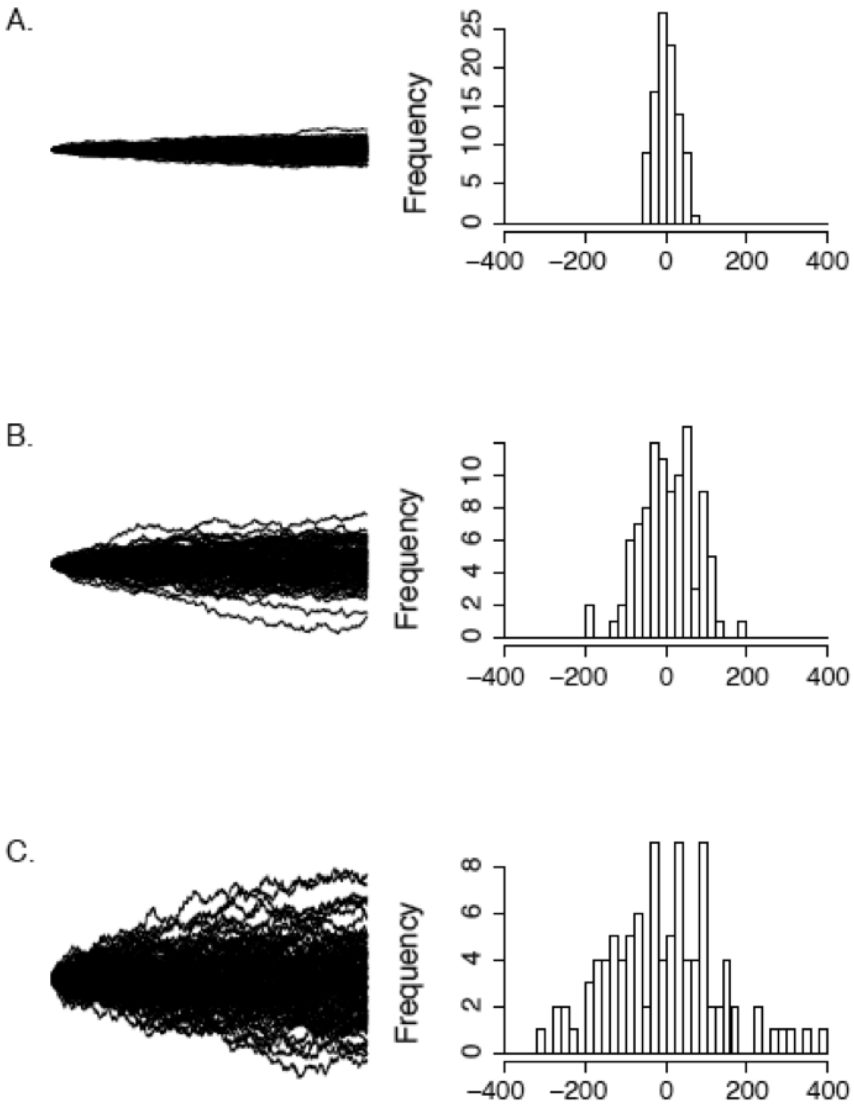
A.
Frequency
0 5 10 15 20 25
−400 −200 0 200 400
B.
Frequency
0 2 4 6 8 10
−400 −200 0 200 400
C.
Frequency
0 2 4 6 8
−400 −200 0 200 400

Figure 3:

If we let $\bar{z}(t)$ be the value of our character at time t, then we can derive three main properties of Brownian motion. I will list all three, then explain each in turn.

1. $E[\bar{z}(t)] = \bar{z}(0)$
2. Each successive interval of the "walk" is independent
3. $\bar{z}(t) \sim N(\bar{z}(0), \sigma^2 t)$

First, $E[\bar{z}(t)] = \bar{z}(0)$. This means that the expected value of the character at any time t is equal to the value of the character at time zero. Here the expected value refers to the mean of $\bar{z}(t)$ over many replicates. The intuitive meaning of this equation is that Brownian motion has no "trends," and wanders equally in both positive and negative directions. If you take the mean of a large number of simulations of Brownian motion over any time interval, you will likely get a value close to $\bar{z}(0)$; as you increase the sample size, this mean will tend to get closer and closer to $\bar{z}(0)$.

Second, each successive interval of the "walk" is independent. Brownian motion is a process in continuous time, and so time does not have discrete "steps." However, if you sample the process at time t, and then again at time $t + \Delta t$, the change that occurs over these two intervals will be independent of one another. This is true of any two non-overlapping intervals sampled from a Brownian walk. It is worth noting that only the changes are independent, and that the value of the walk at time $t + \Delta t$ – which we can write as $\bar{z}(t + \Delta t)$ - is not independent of the value of the walk at time t, $\bar{z}(t)$. But the differences between successive steps [e.g. $\bar{z}(t) - \bar{z}(0)$ and $\bar{z}(t + \Delta t) - \bar{z}(t)$] are independent of each other and of $\bar{z}(0)$.

Finally, $\bar{z}(t) \sim N(\bar{z}(0), \sigma^2 t)$.That is, the value of $\bar{z}(t)$ is drawn from a normal distribution with mean $\bar{z}(0)$ and variance $\sigma^2 t$. As we noted above, the parameter σ^2 is important for Brownian motion models, as it describes the rate at which the process wanders through trait space. The overall variance of the process is that rate times the amount of time that has elapsed.

Section 3.3: Deriving Brownian Motion using Quantitative Genetics

Section 3.3a: Brownian motion under genetic drift

The simplest way to obtain Brownian evolution of characters is when evolutionary change is neutral, with traits changing only due to genetic drift. (e.g. Lande 1976). To show this, we will create a simple model. We will assume that a character is influenced by many genes, each of small effect, and that the value of the character does not affect fitness. Finally, we assume that mutations are random and have small effects on the character, as specified below. These assumptions probably seem unrealistic, especially if you are thinking of a trait like the body

size of a lizard! But we will see later that we can also derive Brownian motion under other models, some of which involve selection.

We again consider the mean value of this trait, $\bar{z}$, in a population with a variance effective population size of N_e. Variance effective population size is the effective population size of a model population with random mating, no substructure, and constant population size that would have quantitative genetic properties equal to our actual population. All of this is a bit beyond the scope of this book (but see Templeton 2006). But writing N_e instead of N allows us to develop the model without worrying about all of the extra assumptions we would have to make about how individuals mate and how populations are distributed over time and space.

Under this model, since there is no selection, the phenotypic character will change due only to mutations and genetic drift. We can model this process in a number of ways, but the simplest uses an infinite alleles model. Under this model, mutations occur randomly and have random phenotypic effects – we can say that mutations are drawn at random from a distribution with mean 0 and mutational variance σ_m^2. This model assumes that the number of alleles is so large that there is effectively no chance of mutations happening to the same allele more than once. The alleles in the population then change in frequency through time due to genetic drift. Drift and mutation together, then, determine the dynamics of the mean trait through time.

If we were to simulate this infinite alleles model many times, we would have a set of evolved populations. These populations would, on average, have the same mean trait value, but would differ from each other. Let's try to derive how, exactly, these populations will differ.

If we consider a population evolving under this model, it is not difficult to show that the expected population phenotype after any amount of time is equal to the starting phenotype. This is because the phenotypes don't matter for survival or reproduction, and mutations are assumed to be symmetrical. Thus,

(eq. 3.1)

$$E[\bar{z}(t)] = \bar{z}(0)$$

Note that this equation already matches the first property of Brownian motion.

Next, we need to also consider the variance of these mean phenotypes, which we will call the between-population phenotypic variance (σ_B^2). Importantly, this is the same quantity we earlier described as the "variance" of traits over time – that is, the variance of mean trait values across many independent "runs" of evolutionary change over a certain time period. To calculate this quantity, we need to consider variation within our model populations. Because of our simplifying assumptions, we only need focus on additive genetic variance within each population at some time t, which we can denote as σ_A^2 (see Lynch and Walsh

1998). Additive genetic variation in a population will change over time due to genetic drift (which tends to decrease σ_A^2) and mutational input (which tends to increase σ_A^2). We can model the expected value of σ_A^2 from one generation to the next as (Clayton and Robertson 1955; Lande 1979, 1980).

(eq. 3.2)

$$E[\sigma_A^2(t+1)] = (1 - \frac{1}{2N_e})E[\sigma_A^2(t)] + \sigma_m^2$$

where t is the elapsed time in generations, N_e is the effective population size, and σ_m^2 is the mutational variance. You can see from this equation that additive genetic variance at time $t+1$ depends on inheritance (σ_A^2 in generation $t+1$ depends on σ_A^2 in generation t), genetic drift (σ_A^2 decreases each generation by a factor that depends on effective population size, N_e), and mutation (σ_A^2 increases by σ_m^2 each generation).

If we assume that we know the starting value at time 0, $\sigma_A^2(0)$, we can calculate the expected additive genetic variance at any time t as:

(eq. 3.3)

$$E[\sigma_A^2(t)] = (1 - \frac{1}{2N_e})^t [\sigma_A^2(0) - 2N_e\sigma_m^2] + 2N_e\sigma_m^2$$

Note that the first term in the above equation, $(1 - \frac{1}{2N_e})^t$, goes to zero as t becomes large. This means that additive genetic variation in the evolving populations will eventually reach an equilibrium between genetic drift and new mutations, so that additive genetic variation stops changing from one generation to the next. We can find this equilibrium by taking the limit of eq. 3.3 as t becomes large.

(eq. 3.4)

$$\lim_{t\to\infty} E[\sigma_A^2(t)] = 2N_e\sigma_m^2$$

Thus the equilibrium genetic variance depends on both population size and mutational input.

We can now derive the between-population phenotypic variance at time t, $\sigma_B^2(t)$. We will assume that σ_A^2 is at equilibrium and thus constant (equation 3.4). Mean trait values in independently evolving populations will diverge from one another. After some time period t has elapsed, that the expected among-population variance will be (from Lande 1976):

(eq. 3.5)

$$\sigma_B^2(t) = \frac{t\sigma_A^2}{N_e}$$

Substituting the equilibrium value of from equation 3.4 into equation 3.5 gives (Lande 1979, 1980):

(eq. 3.6)

$$\sigma_B^2(t) = \frac{t\sigma_A^2}{N_e} = \frac{t \cdot 2N_e\sigma_m^2}{N_e} = 2t\sigma_m^2$$

Notice that for this model, the amount of variation among populations depends only on the rate of mutational input, and is independent of both the starting state of the populations and their effective population size. This model predicts, then, that long-term rates of evolution are dominated by the supply of new mutations to a population.

Lynch and Hill (1986) show that equation 3.6 is a general result that holds under a range of models, even those that include dominance, linkage, nonrandom mating, and other processes. Equation 3.6 is somewhat useful, but we cannot often measure the mutational variance σ_m^2 for any natural populations (but see Turelli 1984). To address this, we can consider the expected heritability for the infinite alleles model at mutational equilibrium. Heritability describes the proportion of total genetic variation within a population (σ_w^2) that is due to additive genetic effects (σ_a^2): $h^2 = \frac{\sigma_A^2}{\sigma_w^2}$. Substituting equation 3.4, we find that:

(eq. 3.7)

$$h^2 = \frac{2N_e\sigma_m^2}{\sigma_w^2}$$

So that:

(eq. 3.8)

$$\sigma_m^2 = \frac{h^2\sigma_w^2}{2N_e}$$

Here, h^2 is heritability, N_e the effective population size, and σ_w^2 the within-population phenotypic variance, which differs from σ_A^2 because it includes all sources of variation within populations, including both non-additive genetic effects and environmental effects. Substituting this expression for σ_w^2 into equation 3.6, we have:

(eq. 3.9)

$$\sigma_B^2(t) = 2\sigma_m^2 t = \frac{h^2\sigma_w^2 t}{N_e}$$

So, after some time interval t, the mean phenotype of a population has an expected value equal to the starting value, and a variance of $\frac{h^2\sigma_w^2 t}{N_e}$.

To derive this result, we had to make particular assumptions about normality of new mutations that might seem quite unrealistic. It is worth noting that if phenotypes are affected by enough mutations, the central limit theorem guarantees that the distribution of phenotypes within populations will be normal – no matter what the underlying distribution of those mutations might be. We also had to assume that traits are neutral, a more dubious assumption that we relax below.

Note, finally, that this quantitative genetics model predicts that traits will evolve under a Brownian motion model. Thus, our quantitative genetics model has the same statistical properties of Brownian motion. We only need to match the parameters: $\Theta = \bar{z}(0)$, and $\sigma^2 = h^2\sigma_w^2/N_e$. In some cases in the literature, the magnitude of trait change is expressed in within-population phenotypic standard deviations, $\sqrt{\sigma_w^2}$, per generation (Estes and Arnold 2007; e.g. Harmon et al. 2010). In that case, since dividing a random normal deviate by x is equivalent to dividing its variance by x^2, we have $\sigma^2 = h^2/N_e$.

Section 3.3b: Brownian motion under selection

We have shown that it is possible to relate a Brownian motion model directly to a quantitative genetics model of drift. In fact, some authors equate the two. However, it is important to remember that the two are not the same thing. More specifically, an observation that a trait is evolving as expected under Brownian motion is not equivalent to saying that that trait is not under selection. This is because characters can also evolve as a Brownian walk even if there is strong selection – as long as selection acts in particular ways that maintain the properties of the Brownian motion model. For example, if the direction and magnitude of selection is random from one generation to the next, then evolution of the character will still follow a Brownian motion model.

In general, the path followed by population mean trait values under mutation, selection, and drift depend on the particular way in which these processes occur. A variety of such models are considered by Hansen and Martins (1996). They identify three very different models that include selection where mean traits still evolve under an approximately Brownian model. Here I present univariate versions of the Hansen-Martins models, for simplicity; consult the original paper for multivariate versions. Note that all of these models require that the strength of selection is relatively weak, or else genetic variation of the character will be depleted by selection over time and the dynamics of trait evolution will change.

One model assumes that populations evolve due to directional selection, but the strength and direction of selection varies randomly from one generation to the next. We model selection each generation as being drawn from a normal distribution with mean 0 and variance σ_s^2. Similar to our drift model, populations will again evolve under Brownian motion. However, in this case the Brownian motion parameters have a different interpretation:

(eq. 3.10)

$$\sigma_B^2 = (\frac{h^2\sigma_W^2}{N_e} + \sigma_s^2)t$$

In the particular case where variation in selection is much greater than variation due to drift, then:

(eq. 3.11)

$$\sigma_B^2\sigma_s^2$$

That is, the drift rate when selection is (on average) much stronger than drift is completely dominated by the selection term. This is not that far fetched, as many studies have shown selection in the wild that is both stronger than drift and commonly changing in both direction and magnitude from one generation to the next.

In a second model, Hansen and Martins (1996) consider a population subject to strong stabilizing selection for a particular optimal value, but where the position of the optimum itself changes randomly according to a Brownian motion process. In this case, population means can again be described by Brownian motion, but now the rate parameter reflects movement of the optimum rather than the action of mutation and drift. Specifically, if we describe movement of the optimum by a Brownian rate parameter σ_E^2, then:

(eq. 3.12)

$$\sigma_B^2\sigma_E^2$$

To obtain this result we must assume that the strength of stabilizing selection is not very weak (at least on the order of $1/t_{ij}$ where t_{ij} is the number of generations separating pairs of populations; Hansen and Martins 1996). Again in this case, the rate of the random walk is totally determined by the action of selection rather than drift.

Finally, Hansen and Martins (1996) consider the situation where populations evolve following a trend. In this case, we get evolution that is different from Brownian motion, but shares some key attributes. Consider a population under constant directional selection, s, so that:

(eq. 3.13)

$$E[\bar{z}(t+1)] = \bar{z}(t) + h^2 s$$

The variance among populations due to genetic drift after a single generation is then:

(eq. 3.14)

$$\sigma_B^2 = \frac{h^2 \sigma_w^2}{N_e}$$

Over some longer period of time, traits will evolve so that they have expected mean trait value that is normal with mean:

(eq. 3.15)

$$E[\bar{z}(t)] = t \cdot (h^2 s)$$

With comparative methods, we are often considering a set of species and their traits in the present day, in which case they will all have experienced the same amount of evolutionary time (t) and have the same expected trait value.

We can also calculate variance among species as:

(eq. 3.16)

$$\sigma_B^2(t) = \frac{h^2 \sigma_w^2 t}{N_e}$$

Note that the variance of this process is exactly identical to the variance among populations in a pure drift model (equation 3.9). Selection only changes the expectation for the species mean (of course, we assume that variation within populations and heritability are constant, which will only be true if selection is quite weak). In fact, equations 3.14 and 3.16 are exactly the same as what we would expect under a pure-drift model in the same population, but starting with a trait value equal to $\Theta = t \cdot (h^2 s)$. That is, from the perspective of data only on living species, these two pure drift and linear selection models are statistically indistinguishable. The implications of this are striking: we can never find evidence for trends in evolution studying only living species.

In summary, we can describe three very different ways that traits might evolve under Brownian motion – pure drift, randomly varying selection, and varying stabilizing selection – and one model, constant directional selection, which creates patterns among extant species that are indistinguishable from Brownian motion. There are certainly more such models, with a variety of assumptions. You might notice that none of these "Brownian" models are particularly detailed,

especially for modeling evolution over long time scales. It is hard to imagine a case where a trait might be influenced only by random mutations of small effect over many alleles, or where selection would act in a truly random way from one generation to the next for millions of years. However, there are tremendous statistical benefits to using Brownian models for comparative analyses. Many of the results derived in this book, for example, are simple under Brownian motion but much more complex and different under other models.

Section 3.4: Brownian motion on a phylogenetic tree

We can use the basic properties of Brownian motion model to figure out what will happen when characters evolve under this model on the branches of a phylogenetic tree. First, consider evolution along a single branch with length t_1 (Figure 3.4A). In this case, we can model simple Brownian motion over time t_1 and denote the starting value as $\bar{z}(0)$. If we evolve with some rate parameter σ^2, then:

(eq. 3.17)

$$E[\bar{z}(t)] \sim N(\bar{z}(0), \sigma^2 t_1)$$

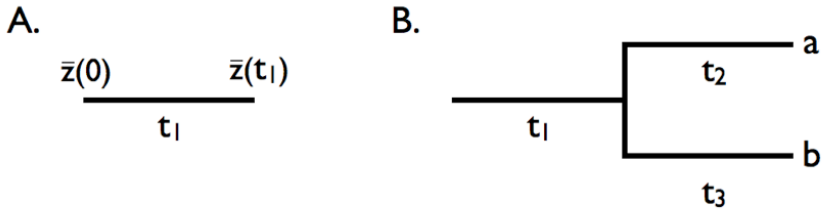

Figure 4:

Figure 3.4. Brownian motion on a simple tree. A. Evolution in a single lineage over time period t_1. B. Evolution on a phylogenetic tree relating species a and b, with branch lengths as given by t_1, t_2, and t_3.

Now consider a small section of a phylogenetic tree including two species and an ancestral stem branch (Figure 3.4B). Assume a character evolves on that tree under Brownian motion, again with starting value $\bar{z}(0)$ and rate parameter σ^2. First consider species a. The mean trait in that species $\bar{x_a}$ evolves under Brownian motion from the ancestor to species a over a total time of $t1 + t2$. Thus,

(eq. 3.18)

$$\bar{x_a} \sim N[\bar{z}(0), \sigma^2(t_1 + t_2)]$$

Similarly for species b, over a total time of $t1 + t3$

(eq. 3.19)

$$\bar{x_b} \sim N[\bar{z}(0), \sigma^2(t_1 + t_3)]$$

However, $\bar{x_a}$ and $\bar{x_b}$ are not independent of each other. Instead, the two species share one branch in common (along branch 1). Each tip trait value can be thought of as the sum of two normal deviates, one (from branch 1) that is shared between the two species and one that is unique (branch 2 for species a and branch 3 for species b). In this case, mean trait values $\bar{x_a}$ and $\bar{x_b}$ will share similarity due to their shared evolutionary history. We can describe this similarity by calculating the covariance between the traits of species a and b. We note that:

(eq. 3.20)

$$\begin{aligned} \bar{x_a} &= \Delta\bar{x_1} + \Delta\bar{x_2} \\ \bar{x_b} &= \Delta\bar{x_1} + \Delta\bar{x_3} \end{aligned}$$

Where $\Delta\bar{x_1}$, $\Delta\bar{x_2}$, and $\Delta\bar{x_3}$ represent evolution along the three branches in the tree, are all normally distributed with mean zero and variances $\sigma 2t1$, $\sigma 2t2$, and $\sigma 2t3$, respectively. $\bar{x_a}$ and $\bar{x_b}$ are sums of normal random variables and are themselves normal. The covariance of these two terms is simply the variance of their shared term:

(eq. 3.21)

$$cov(\bar{x_a}, \bar{x_b}) = var(\Delta\bar{x_1}) = \sigma^2 t_1$$

In fact, the trait values for the two species are drawn from a multivariate normal distribution. Each trait has the same expected value, Θ, and the two traits have a variance-covariance matrix:

(eq. 3.22)

$$\begin{bmatrix} \sigma^2(t_1 + t_2) & \sigma^2 t_1 \\ \sigma^2 t_1 & \sigma^2(t_1 + t_3) \end{bmatrix} = \sigma^2 \begin{bmatrix} t_1 + t_2 & t_1 \\ t_1 & t_1 + t_3 \end{bmatrix}$$

The matrix on the right side of equation 3.22 is commonly encountered in comparative biology, and will come up again in this book. We will call this matrix the phylogenetic variance-covariance matrix, $\mathbf{C}$. This matrix has a special structure. For phylogenetic trees with n species, this is an nxn matrix, with each

row and column corresponding to one of the n taxa in the tree. Along the diagonal are the total distances of each taxon from the root of the tree, while the off-diagonal elements are the total branch lengths shared by particular pairs of taxa. For example, $\mathbf{C}(1,2)$ and $\mathbf{C}(2,1)$ – which are equal because the matrix $\mathbf{C}$ is always symmetric – is the shared phylogenetic path length between the species in the first row – here, species a - and the species in the second row – here, species b. Under Brownian motion, these shared path lengths are proportional to the phylogenetic covariances of trait values. A full example of a phylogenetic variance-covariance matrix for a small tree is shown in Figure 3.5. This multivariate normal distribution completely describes the expected statistical distribution of traits on the tips of a phylogenetic tree if the traits evolve according to a Brownian motion model.

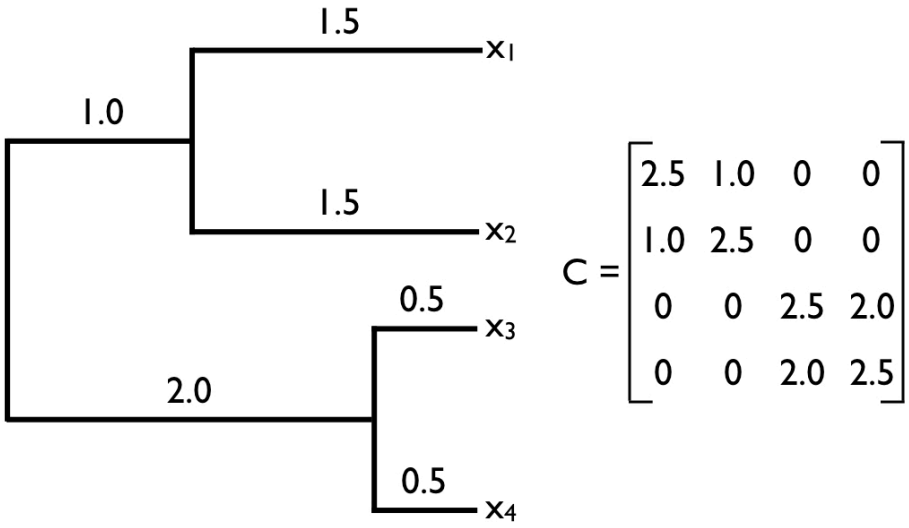

Figure 5:

Figure 3.5. Example of a phylogenetic tree (left) and its associated phylogenetic variance-covariance matrix $\mathbf{C}$ (right).

Section 3.5: Multivariate Brownian motion

The Brownian motion model we described above was for a single character. However, we often want to consider more than one character at once. This

requires the use of multivariate models. The situation is more complex than the univariate case – but not much! In this section I will derive the expectation for a set of (potentially correlated) traits evolving together under a multivariate Brownian motion model.

Character values across species can covary because of phylogenetic relationships, because different characters tend to evolve together, or both. Fortunately, we can generalize the model described above to deal with both of these types of covariation. To do this, we must combine two variance-covariance matrices. The first one, $\mathbf{C}$, we have already seen; it describes the variances and covariances across *species* for single traits due to shared evolutionary history along the branches of a phylogentic tree. The second variance-covariance matrix, which we can call $\mathbf{R}$, describes the variances and covariances across *traits* due to their tendencies to evolve together. For example, if a species of lizard gets larger due to the action of natural selection, then many of its other traits, like head and limb size, will get larger too due to allometry. The diagonal entries of the matrix $\mathbf{R}$ will provide our estimates of σ_i^2, the net rate of evolution, for each trait, while off-diagonal elements represent evolutionary covariances between pairs of traits. We will denote number of species as n and the number of traits as m, so that $\mathbf{C}$ is $n \times n$ and $\mathbf{R}$ is $m \times m$.

Our multivariate model of evolution has parameters that can be described by an $m \times 1$ vector, $\mathbf{a}$, containing the starting values for each trait – $\bar{z_1}(0)$, $\bar{z_2}(0)$, and so on, up to $\bar{z_m}(0)$, and an mxm matrix, $\mathbf{R}$, described above. This model has m parameters for $\mathbf{a}$ and $m \cdot (m+1)/2$ parameters for $\mathbf{R}$, for a total of $m \cdot (m+3)/2$ parameters.

Under our multivariate Brownian motion model, the joint distribution of all traits across all species still follows a multivariate normal distribution. We find the variance-covariance matrix that describes all characters across all species by combining the two matrices $\mathbf{R}$ and $\mathbf{C}$ into a single large matrix using the Kroeneker product:

(eq. 3.23)

$$\mathbf{V} = \mathbf{R} \otimes \mathbf{C}$$

This matrix $\mathbf{V}$ is $n \cdot m \times n \cdot m$, and describes the variances and covariances of all traits across all species.

We can return to our example of evolution along a single branch (Figure 3.4a). Imagine that we have two characters that are evolving under a multivariate Brownian motion model. We state the parameters of the model as:

(eq. 3.24)

$$\mathbf{a} = \begin{bmatrix} \bar{z_1}(0) \\ \bar{z_2}(0) \end{bmatrix}$$
$$\mathbf{R} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

For a single branch, $\mathbf{C} = [t_1]$, so:

(eq. 3.25)

$$\mathbf{V} = \mathbf{R} \otimes \mathbf{C} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \otimes [t_1] = \begin{bmatrix} \sigma_1^2 t_1 & \sigma_{12} t_1 \\ \sigma_{12} t_1 & \sigma_2^2 t_1 \end{bmatrix}$$

The two traits follow a multivariate normal distribution with mean a and variance-covariance matrix $\mathbf{V}$.

For the simple tree in figure 3.4b,

(eq. 3.26)

$$\mathbf{V} = \mathbf{R} \otimes \mathbf{C} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \otimes \begin{bmatrix} t_1 + t_2 & t_1 \\ t_1 & t_1 + t_3 \end{bmatrix}$$
$$= \begin{bmatrix} \sigma_1^2(t_1 + t_2) & \sigma_{12}(t_1 + t_2) & \sigma_1^2 t_1 & \sigma_{12} t_1 \\ \sigma_{12}(t_1 + t_2) & \sigma_2^2(t_1 + t_2) & \sigma_{12} t_1 & \sigma_2^2 t_1 \\ \sigma_1^2 t_1 & \sigma_{12} t_1 & \sigma_1^2(t_1 + t_3) & \sigma_{12}(t_1 + t_3) \\ \sigma_{12} t_1 & \sigma_2^2 t_1 & \sigma_{12}(t_1 + t_3) & \sigma_2^2(t_1 + t_3) \end{bmatrix}$$

Thus, the four trait values (two traits for two species) are drawn from a multivariate normal distribution with mean $a = [\bar{z_1}(0), \bar{z_1}(0), \bar{z_2}(0), \bar{z_2}(0)]$ and the variance-covariance matrix shown above.

Both univariate and multivariate Brownian motion models result in traits that follow multivariate normal distributions. This is statistically convenient, and in part explains the popularity of Brownian models in comparative biology.

Section 3.6: Simulating Brownian motion on trees

To simulate Brownian motion evolution on trees, we use the three properties of the model described above. For each branch on the tree, we can draw from a normal distribution (for a single trait) or a multivariate normal distribution (for more than one trait) to determine the evolution that occurs on that branch. We can then add these evolutionary changes together to obtain character states at every node and tip of the tree.

I will illustrate one such simulation for the simple tree depicted in figure 3.4b. We first set the ancestral character state to be , which will then be the expected

value for all the nodes and tips in the tree. This tree has three branches, so we draw three values from normal distributions. These normal distributions have variances that are given by the rate of evolution and the branch length of the tree, as stated in equation 3.1. Note that we are modeling changes on these branches, so even if $\bar{z_1}(0)0$ the values for changes on branches are drawn from a distribution with a mean of zero. In the case of the tree in Figure 3.1, $x1 \sim N(0, \sigma^2 t_1)$. Similarly, $x2 \sim N(0, \sigma^2 t_2)$ and $x3 \sim N(0, \sigma^2 t_3)$. If I set $\sigma^2 = 1$ for the purposes of this example, I might obtain $x_1 = -1.6$, $x_2 = 0.1$, and $x_3 = -0.3$. These values represent the evolutionary changes that occur along branches in the simulation. To calculate trait values for species, we add: $x_a = +x_1 + x_2 = 0 - 1.6 + 0.1 = -1.5$, and $x_b = +x_1 + x_3 = 0 - 1.6 + -0.3 = -1.9$.

This simulation algorithm works fine but is actually more complicated than it needs to be, especially for large trees. We already know that xA and xB come from a multivariate normal distribution with known mean vector and variance-covariance matrix. We can simply draw a vector from this distribution, and our tip values will have exactly the same statistical properties as if they were simulated on a phylogenetic tree. These two methods for simulating character evolution on trees are exactly equivalent to one another.

In this chapter, we consider Brownian motion, and first connected that process to a model of genetic drift for traits that have no effect on fitness. However, Brownian motion can result from a variety of other models, some of which include natural selection. For example, traits will follow Brownian motion under selection is if the strength and direction of selection varies randomly through time. As the time intervals between samples becomes large relative to the frequency of selection, then evolution will follow a Brownian model.

There is a general feature of models that evolve in a Brownian way: they involve the action of a large number of very small "forces" pushing on characters. No matter the particular distribution of these small effects or even what causes them, if you add together enough of them you will obtain a normal distribution of outcomes and, sometimes, be able to model this process using Brownian motion. The main restriction might be the unbounded nature of Brownian motion – species are expected to become more and more different through time, without any limit, which must be unrealistic over very long time scales.

In summary, Brownian motion is mathematically tractable, and has convenient statistical properties. There are also some circumstances under which one would expect traits to evolve under a Brownian model. However, as we will see later in the book, one should view Brownian motion as an assumption that might not hold for real data sets.

Chapter 3 References

Clayton, G., and A. Robertson. 1955. Mutation and quantitative variation. Am. Nat. 89:151–158.

Estes, S., and S. J. Arnold. 2007. Resolving the paradox of stasis: Models with stabilizing selection explain evolutionary divergence on all timescales. Am. Nat. 169:227–244.

Hansen, T. F., and E. P. Martins. 1996. TRANSLATING BETWEEN MICROEVOLUTIONARY PROCESS AND MACROEVOLUTIONARY PATTERNS: THE CORRELATION STRUCTURE OF INTERSPECIFIC DATA. Evolution 50:1404–1417.

Harmon, L. J., J. B. Losos, T. Jonathan Davies, R. G. Gillespie, J. L. Gittleman, W. Bryan Jennings, K. H. Kozak, M. A. McPeek, F. Moreno-Roark, T. J. Near, and Others. 2010. Early bursts of body size and shape evolution are rare in comparative data. Evolution 64:2385–2396.

Hedges, B. S., and S. Kumar. 2009. The timetree of life. Oxford University Press, Oxford.

Lande, R. 1976. Natural selection and random genetic drift in phenotypic evolution. Evolution 30:314–334.

Lande, R. 1979. Quantitative genetic analysis of multivariate evolution, applied to brain:body size allometry. Evolution 33:402–416.

Lande, R. 1980. Sexual dimorphism, sexual selection, and adaptation in polygenic characters. Evolution 34:292–305.

Lynch, M., and W. G. Hill. 1986. Phenotypic evolution by neutral mutation. Evolution 40:915–935.

Lynch, M., and B. Walsh. 1998. Genetics and analysis of quantitative traits. Sinauer Sunderland, MA.

Templeton, A. R. 2006. Population genetics and microevolutionary theory. John Wiley & Sons.

Turelli, M. 1984. Heritable genetic variation via mutation-selection balance: Lerch's zeta meets the abdominal bristle. Theor. Popul. Biol. 25:138–193.

Chapter 4: Fitting Brownian Motion Models to Single Characters

Section 4.1: Introduction

Mammals come in a wide variety of shapes and sizes. Some species are incredibly tiny. For example, the bumblebee bat, weighing in at 2 g, competes for the title of smallest mammal with the slightly lighter (but also slightly longer) Etruscan shrew. Other species are huge, as anyone who has encountered a blue whale knows. Body size is important as a biological variable because it predicts so many other aspect of an animal's life, from the physiology of heat exchange to the biomechanics of locomotion. Thus, the evolutionary rate of body size evolution is of great interest among mammalian biologists. Throughout this chapter, I will discuss the evolution of body size (and, eventually, territory size) across different species of mammals. The data I will analyze is taken from Garland (1992).

Sometimes one might be interested in calculating the rate of evolution of a particular character like body size in a certain clade, say, mammals1. You have a phylogenetic tree with branch lengths that are proportional to time, and data on the phenotypes of species on the tips of that tree. It is usually a good idea to log-transform your data if they involve a measurement from a living thing (see Box 4.1, below). If we assume that the character has been evolving under a Brownian motion model, we have two parameters to estimate: $\bar{z}(0)$, the starting value for the Brownian motion model – equivalent to the ancestral state of the character at the root of the tree – and σ^2, the diffusion rate of the character. It is this latter parameter that is commonly considered as the rate of evolution for comparative approaches.

Box 4.1: Biology under the log

One general rule for continuous traits in biology is to carry out a log-transformation (usually natural log, base e) of your data before undergoing any analysis. This also applies to comparative data. There are two main reasons for this, one statistical and the other biological. The statistical reason is that many methods assume that variables follow normal distributions. One can observe that, in general, biological variables have a distribution that is skewed to the right. A log-transformation will often result in trait distributions that are closer to normal. But why is this the case? The answer is related to the biological reason for log-transformation. When you log transform a variable, the new scale for that variable is a ratio scale, so that a certain differences between points reflects a constant ratio of the two numbers represented by the points. So, for example, if any two numbers are separated by 0.693 units on a

natural log scale, one will be exactly two times the other. Ratio scales make sense for living things because it is usually percentage changes rather than absolute changes that matter. For example, a change in body size of 1 mm might matter a lot for a termite, but be irrelevant for an elephant; whereas a change in body size of 50% might be expected to matter for them both.

Section 4.2: Estimating rates using independent contrasts

The information required to estimate evolutionary rates is efficiently summarized in the early (but still useful) phylogenetic comparative method of independent contrasts (Felsenstein 1985). Independent contrasts summarize the amount of character change across each node in the tree, and can be used to estimate the rate of character change across a phylogeny. There is also a simple mathematical relationship between contrasts and maximum-likelihood rate estimates that I will discuss below.

We can understand the basic idea behind independent contrasts if we think about the branches in the phylogenetic tree as the historical "pathways" of evolution. Each branch on the tree represents a lineage that was alive at some time in the history of the Earth, and during that time experienced some amount of evolutionary change. We can imagine trying to measure that change initially by comparing sister taxa. We can compare the trait values of the two sister taxa by finding the difference in their trait values, and then compare that to the total amount of time they have had to evolve that difference. By doing this for all sister taxa in the tree, we will get an estimate of the average rate of character evolution (4.1A). But what about deeper nodes in the tree? We could use other non-sister species pairs, but then we would be counting some branches in the tree of life more than once (Figure 4.1B). Instead, we use a "pruning algorithm," [Felsenstein1985-bt, Felsenstein2004-eo] chopping off pairs of sister taxa to create a smaller tree (Figure 4.1C). Eventually, all of the nodes in the tree will be trimmed off – and the algorithm will finish. Independent contrasts provides a way to generalize the approach of comparing sister taxa so that we can quantify the rate of evolution throughout the whole tree.

A more precise algorithm describing how phylogenetic independent contrasts are calculated is provided in Box 4.2, below (Felsenstein 1985). Each contrast can be described as an estimate of the direction and amount of evolutionary change across the nodes in the tree. PICs are calculated from the tips of the tree towards the root, as differences between trait values at the tips of the tree and/or calculated average values at internal nodes. The differences themselves are sometimes called "raw contrasts" (Felsenstein 1985). These raw contrasts will all be statistically independent of each other under a wide range of evolutionary models – in fact, as long as each lineage in a phylogenetic tree evolves independently of every other lineage, regardless of the evolutionary model, the

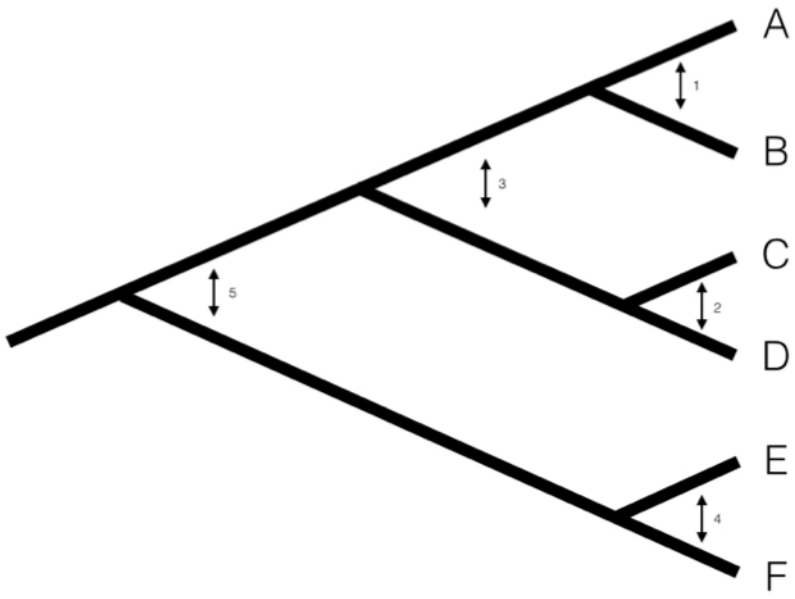

Figure 1: Figure 4.1. Pruning algorithm that can be used to identify five independent contrasts for a tree with six species (from Felsenstein 1985). The numbered order in this figure is only one of two possibilities that work; one can also prune the tree in the order 1, 2, 4, 3, 5 and get identical results.

raw contrasts will be independent of each other. However, people almost never use raw contrasts because they are not identically distributed; each raw contrast has a different expected distribution that depends on the model of evolution and the branch lengths of the tree. Felsenstein (1985) divided the raw contrasts by their expected standard deviation under a Brownian motion, resulting in standardized contrasts. These standardized contrasts are, under a BM model, both independent and identically distributed, and can be used in a variety of statistical tests.

Box 4.2: Algorithm for PICs

One can calculate PICs using the algorithm from Felsenstein (1985). I reproduce this algorithm below. Keep in mind that this is an iterative algorithm – you repeat the five steps below once for each contrast, or $n-1$ times over the whole tree (see Figure 4.1C as an example).

1. Find two tips on the phylogeny that are adjacent (say nodes i and j) and have a common ancestor, say node k. Note that the choice of which node is i and which is j is arbitrary. We will have to account for this "arbitrary direction" property of PICs in any analyses where we use them to do statistics!

2. Compute the contrast, the difference between their two tip values:

   (eq. 4.1)

$$c_{ij} = x_i - x_j$$

Under a Brownian motion model, c_{ij} has expectation zero and variance proportional to $v_i + v_j$.

3. Calculate the standardized contrast by dividing the raw contrast by its variance

   (eq. 4.2)

$$s_{ij} = \frac{c_{ij}}{v_i + v_j} = \frac{x_i - x_j}{v_i + v_j}$$

Under a Brownian motion model, this contrast has mean zero and variance equal to the Brownian motion rate parameter σ^2.

4. Remove the two tips from the tree, leaving behind only the ancestor k, which now becomes a tip. Assign it the character value:

(eq. 4.3)

$$x_k = \frac{(1/v_i)x_i + (1/v_j)x_j}{1/v_1 + 1/v_j}$$

It is worth noting that x_k is a weighted average of x_i and x_j, but does not represent an ancestral state reconstruction, since the value is only influenced by species that descend directly from that node and not other relatives.

5. Lengthen the branch below node k by increasing its length from v_k to $v_k + v_i v_j/(v_i + v_j)$. This accounts for the error in assigning a value to x_k.

As mentioned above, we can apply the algorithm of independent contrasts to learn something about rates of body size evolution in mammals. We have a phylogenetic tree with branch lengths as well as body mass estimates for 49 species (Figure 4.2). If we ln-transform mass and then apply the method above to our data on mammal body size, we obtain a set of 48 standardized contrasts. A histogram of these contrasts is shown as Figure 4.2.

Note that each contrast is an amount of change $x_i - x_j$ divided by a branch length $v_i + v_j$, which is a measure of time. Thus, PICs from a single trait can be used to estimate , the net rate of evolution under a Brownian model. The PIC estimate of the evolutionary rate is:

(eq. 4.4)

$$\hat{\sigma}^2_{PIC} = \frac{\sum s^2_{ij}}{n-1}$$

Here, the sum is taken over all s_{ij}, the standardized independent contrast across all (i, j) pairs of sister branches in the phylogenetic tree. For a fully bifurcating tree with n tips, there are exactly $n-1$ such pairs. If you are statistically savvy, you might note that this formula looks a bit like a variance. In fact, if we state that the contrasts have a mean of 0 (which they must because Brownian motion has no overall trends), then this is a formula to estimate the variance of the contrasts.

If we calculate the mean sum of squared contrasts for the mammal body mass data, we obtain a rate estimate of $\hat{\sigma}^2_{PIC} = 0.90$.

Section 4.3: Estimating rates using maximum likelihood

We can also estimate the evolutionary rate by finding the maximum-likelihood parameter values for a Brownian motion model fit to our data. Recall that ML

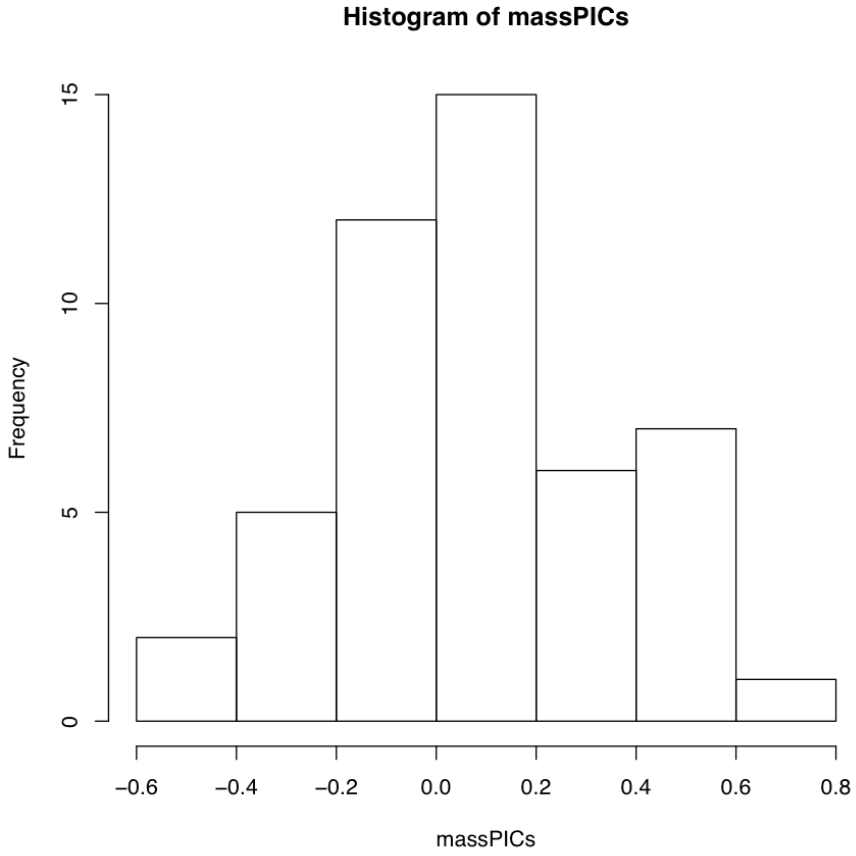

Figure 2: Figure 4.2. Histogram of PICs for ln-transformed mammal body mass on a phylogenetic tree with branch lengths in millions of years (from Garland 1992).

parameter values are those that maximize the likelihood of the data given our model.

Under a Brownian motion model tip character states are drawn from a multivariate normal distribution with a variance-covariance matrix, $\mathbf{C}$, that is calculated based on the branch lengths and topology of the phylogenetic tree (see Chapter 3). We can calculate the likelihood of obtaining the data under our Brownian motion model:

(eq. 4.5)

$$L(\mathbf{x}|\bar{z}(0), \sigma^2, \mathbf{C}) = \frac{e^{-1/2(\mathbf{x}-\bar{z}(0)\mathbf{1})^\intercal(\sigma^2\mathbf{C})^{-1}(\mathbf{x}-\bar{z}(0)\mathbf{1})}}{\sqrt{(2\pi)^n det(\sigma^2\mathbf{C})}}$$

Here, our model parameters are σ^2 and $\bar{z}(0)$, the root trait value. x is an $n \times 1$ vector of trait values for the n tip species in the tree, with species in the same order as $\mathbf{C}$, and $\mathbf{1}$ is an $n \times 1$ column vector of ones. Note that $(\sigma^2\mathbf{C})^{-1}$ is the matrix inverse of the matrix $\sigma^2\mathbf{C}$

As an example, with the mammal data, we can calculate the likelihood for a model with parameter values $\sigma^2 = 1$ and $\bar{z}(0) = 0$ as $L(\mathbf{x}|\bar{z}(0), \sigma^2, \mathbf{C}) = -116.2$.

To find the ML estimates of our model parameters, we need to find the parameter values that maximize that function. One (not very efficient) way to do this is to calculate the likelihood across a wide range of parameter values. One can then visualize the resulting likelihood surface and identify the maximum of the likelihood function. For example, the likelihood surface for the mammal body size data given a Brownian motion model is shown in Figure 4.3. Note that this surface has a peak around $\sigma^2 = 0.09$ and $\bar{z}(0) = 4$. Inspecting the matrix of ML values, we find the highest likelihood (-78.05) at $\sigma^2 = 0.089$ and $\bar{z}(0) = 4.65$.

The calculation described above is inefficient, because we have to calculate likelihoods at a wide range of parameter values that are far from the optimum. To be more efficient, we can use numerical optimization, a branch of computer science that is dedicated to finding efficient algorithms that search for the optima (minima or maxima) of functions. One simple example is based on Newton's method of optimization [as implemented, for example, by the r function nlm()]. We can use this algorithm to quickly find accurate ML estimates2.

Using optimization algorithms we find a ML solution at $\sigma^2 = 0.08804487$ and $\bar{z}(0) = 4.640571$, with $lnL = -78.04942$. Importantly, the solution is located with only 10 likelihood calculations. I have plotted the path through parameter space taken by Newton's method when searching for the optimum in Figure 4.4. Notice two things: first, that the function starts at some point and heads uphill on the likelihood surface until an optimum is found; and second, that this calculation requires many fewer steps (and much less time) than calculating the likelihood for a wide range of parameter values.

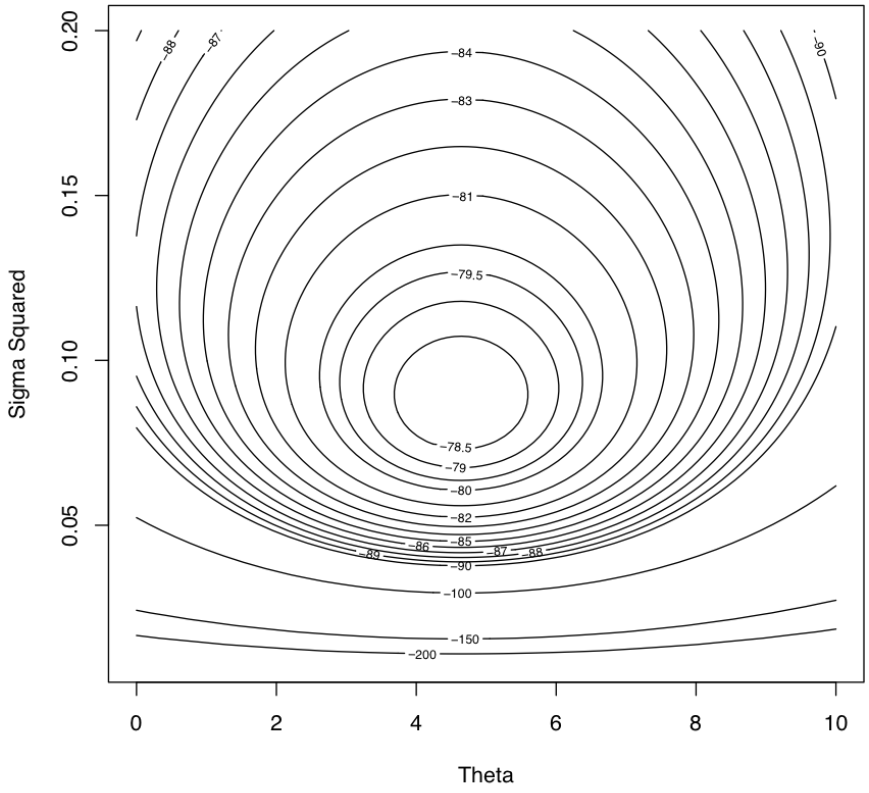

Figure 3: Figure 4.3. Likelihood surface for the evolution of mammalian body mass using the data from Garland (1992).

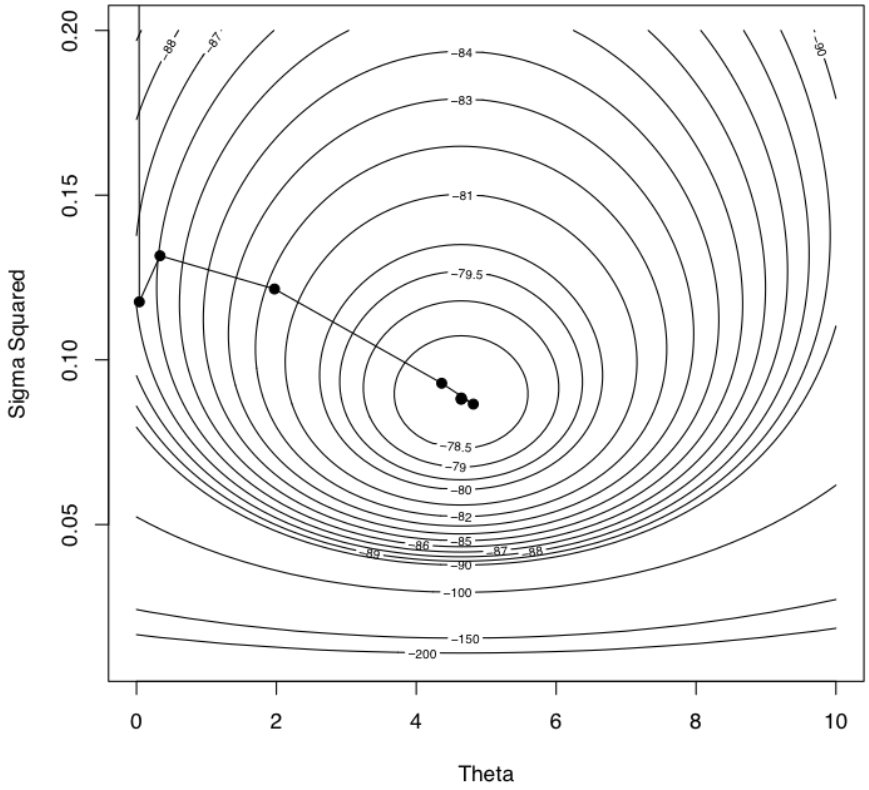

Figure 4: Figure 4.4. Likelihood surface for the evolution of mammalian body mass using the data from Garland (1992). Shown here is the path taken by the optimization algorithm to find the peak of the likelihood surface. The last five steps of this ten-step algorithm are too close together to be seen in this figure.

Using an optimization algorithm also has the added benefit of providing (approximate) confidence intervals for parameter values based on the hessian of the likelihood surface. This approach assumes that the shape of the likelihood surface in the immediate vicinity of the peak can be approximated by a quadratic function, and uses the curvature of that function to approximate the standard errors of parameter values (Burnham and Anderson 2003). If the surface is strongly peaked, the SEs will be small, while if the surface is very broad, the SEs will be large. For example, the likelihood surface around the ML values for mammal body size evolution has a Hessian of:

(eq. 4.6)

$$H = \begin{bmatrix} 314.6 & -0.0026 \\ -0.0026 & 0.99 \end{bmatrix}$$

This gives standard errors of 0.13 (for σ^2) and 0.72 [for $\bar{z}(0)$]. If we assume the error around these estimates is approximately normal, we can create confidence estimates by adding and subtracting twice the standard error. We then obtain 95% CIs of 0.060.11 (for σ^2) and $3.22 - 6.06$ [for $\bar{z}(0)$].

The danger in optimization algorithms is that one can sometimes get stuck on local peaks. More elaborate algorithms repeated for multiple starting points can help solve this problem, but are not needed for simple Brownian motion on a tree as considered here. Numerical optimization is a pressing problem in phylogenetic comparative methods that I will return to later in the book.

In the particular case of fitting Brownian motion to trees, it turns out that even our fast algorithm for optimization was unnecessary. In this case, the maximum-likelihood estimate for each of these two parameters can be calculated analytically.

(eq. 4.7)

$$\hat{\bar{z}}(0) = (\mathbf{1}\mathbf{C}^{-1}\mathbf{1})^{-1}(\mathbf{1}\mathbf{C}^{-1}\mathbf{x})$$

and:

(eq. 4.8)

$$\hat{\sigma}^2_{ML} = \frac{(\mathbf{x} - \hat{\bar{z}}(0)\mathbf{1})\mathbf{C}^{-1}(\mathbf{x} - \hat{\bar{z}}(0)\mathbf{1})}{n}$$

where n is the number of taxa in the tree, $\mathbf{C}$ is the $n \times n$ variance-covariance matrix under Brownian motion for tip characters given the phylogenetic tree, x is an $n \times 1$ vector of trait values for tip species in the tree, $\mathbf{1}$ is an $n \times 1$ column vector of ones, $\hat{\bar{z}}(0)$ is the estimated root state for the character, and $\hat{\sigma}^2_{ML}$ is the estimated net rate of evolution.

Applying this approach to mammal body size, we obtain estimates that are exactly the same as our results from numeric optimization: $\sigma^2 = 0.088$ and $\bar{z}(0) = 4.64$.

Equation (4.8) is biased, and will consistently estimate rates of evolution that are a little too small; an unbiased version based on restricted maximum likelihood (REML) and used by Garland (1992) and others is:

(eq. 4.9)

$$\hat{\sigma}^2_{REML} = \frac{(\mathbf{x} - \hat{\bar{z}}(0)\mathbf{1})\mathbf{C}^{-1}(\mathbf{x} - \hat{\bar{z}}(0)\mathbf{1})}{n-1}$$

This correction changes our estimate of the rate of body size in mammals from $\sigma^2 = 0.088$ to $\sigma^2 = 0.090$. Equation 4.8 is exactly identical to the estimated rate of evolution calculated using the average squared independent contrast, described above; that is, $\hat{\sigma}^2_{PIC} = \hat{\sigma}^2_{REML}$. In fact, PICs are a formulation of a REML model. The "restricted" part of REML refers to the fact that these methods calculate likelihoods based on a transformed set of data where the effect of nuisance parameters has been removed. In this case, the nuisance parameter is the estimated root state $\hat{\bar{z}}(0)$. PICs are a transformation of the original data in which all information about the root state has been removed; our idea of what that root state might be has no effect on calculations using PICs. One can calculate the likelihood for the PIC REML method by assuming all of the standardized PICs are drawn from a normal distribution (eq. 4.5) with mean 0 and variance $\hat{\sigma}^2_{REML}$ (eq. 4.8). Alternatively, one can estimate the variance of the PICs directly, keeping in mind that one must use a mean of zero (eq. 4.4). These two methods give exactly the same results.

For the mammal body size example, we can further explore the difference between REML and ML in terms of statistical confidence intervals using likelihoods based on the contrasts. We assume, again, that the contrasts are all drawn from a normal distribution with mean 0 and unknown variance. If we again use Newton's method for optimization, we find a maximum REML log-likelihood of -10.3 at $\hat{\sigma}^2_{REML} = 0.90$. This returns a $1times1$ matrix for the Hessian with a value of 2957.8, corresponding to a SE of 0.018. This slightly larger SE corresponds to 95% CI for $\hat{\sigma}^2_{REML}$ of 0.050.13.

In the context of comparative methods, REML has two main advantages. First, PICs treat the root state of the tree as a nuisance parameter. We typically have very little information about this root state, so that can be an advantage of the REML approach. Second, PICs are easy to calculate for very large phylogenetic trees because they do not require the construction (or inversion!) of any large variance-covariance matrices. This is important for big phylogenetic trees. Imagine that we had a phylogenetic tree of all vertebrates (~60,000 species) and wanted to calculate the rate of body size evolution. To use standard maximum likelihood, we have to calculate $\mathbf{C}$, a matrix with $60,000 \times 60,000 = 3.6$ billion entries, and invert it to calculate $\mathbf{C}^{-1}$. To calculate PICs, by contrast, we only have to carry out on the order of 120,000 operations. Thankfully, there are

now pruning algorithms to quickly calculate likelihoods for large trees under a variety of different models [see, e.g., FitzJohn (2012), Freckleton (2012), and Ho and Ané (2014)).

Section 4.4: Bayesian approach to evolutionary rates

Finally, we can also use a Bayesian approach to fit Brownian motion models to data and to estimate the rate of evolution. This approach differs from the ML approach in that we will use explicit priors for parameter values, and then run an MCMC to estimate posterior distributions of parameter estimates. To do this, we will modify the basic algorithm for Bayesian MCMC (see Chapter 2) as follows:

1. Sample a set of starting parameter values, σ^2 and $\bar{z}(0)$ from their prior distributions. For this example, we can set our prior distribution as uniform between 0 and 1 for σ^2 and uniform from -1 to +1 for $\bar{z}(0)$.
2. Given the current parameter values, select new proposed parameter values using the proposal density $Q(p'|p)$. For both parameter values, we will use a uniform proposal density with width w_p, so that:

(eq. 4.10)

$$Q(p'|p) \sim U(p - \frac{w_p}{2}, p + \frac{w_p}{2})$$

3. Calculate three ratios:

a. The prior odds ratio. This is the ratio of the probability of drawing the parameter values p and p from the prior. Since our priors are uniform, this is always 1.
b. The proposal density ratio. This is the ratio of probability of proposals going from p to p and the reverse. We have already declared a symmetrical proposal density, so that $Q(p'|p) = Q(p|p')$ and $a_2 = 1$.
c. The likelihood ratio. This is the ratio of probabilities of the data given the two different parameter values. We can calculate these probabilities from equation 4.5 above.

(eq. 4.11)

$$a_3 = \frac{L(p'|D)}{L(p|D)} = \frac{P(D|p')}{P(D|p)}$$

4. Find the product of the prior odds, proposal density ratio, and the likelihood ratio. In this case, both the prior odds and proposal density ratios are 1, so $a = a_3$.

5. Draw a random number x from a uniform distribution between 0 and 1. If $x < a$, accept the proposed value of both parameters; otherwise reject, and retain the current value of the two parameters.

6. Repeat steps 2-5 a large number of times.

Using the mammal body size data, I ran the analysis with uniform priors from (0, 0.5) for σ_2 and from (0, 10) for $\bar{z}(0)$. I used an MCMC with 10,000 generations, discarding the first 1000 as burn-in. Sampling every 10 generations, I obtain parameter estimates of $\sigma_2 = 0.10$ (95% credible interval: 0.0660.15) and $\bar{z}(0) = 3.5$ (95% credible interval: 2.35.3; Figure 4.5).

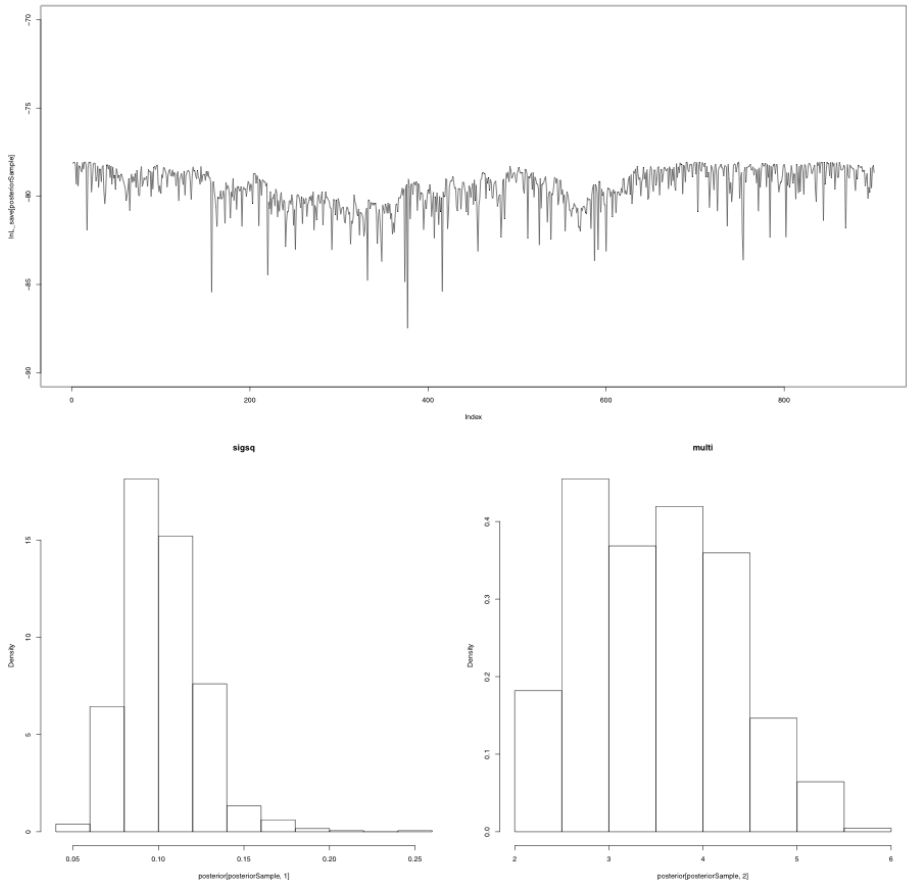

Figure 5: Figure 4.5. Bayesian analysis of body size evolution in mammals. Figure shows the likelihood profile (A) and posterior distributions for model parameters σ_2 (B) and $\bar{z}(0)$ (C).

Note that the parameter estimates from all three approaches (REML, ML, and

Bayesian) were similar. Even the confidence/credible intervals varied a little bit but were of about the same size in all three cases. All of the approaches above are mathematically related and should, in general, return similar results. One might place higher value on the Bayesian credible intervals over confidence intervals from the Hessian of the likelihood surface, for two reasons: the Hessian leads to an estimate of the CI under certain conditions that may or may not be true for your analysis; and second, Bayesian credible intervals reflect overall uncertainty better than ML confidence intervals (see chapter 2).

Section 4.5: Summary

By fitting a Brownian motion model to phylogenetic comparative data, one can estimate the rate of evolution of a single character. In this chapter, I demonstrated three approaches to estimating that rate: PICs, maximum likelihood, and Bayesian MCMC. In the next chapter, we will discuss other models of evolution that can be fit to continuous characters on trees.

Footnotes

1: Throughout this chapter, when I say rate I will mean the Brownian motion rate parameter. This is a little different from "traditional" estimates of evolutionary rate, like those estimated by paleontologists. For example, one might have measurements of trait in a series of fossils representing an evolutionary lineage sampled at different time periods. By calculating the amount of change over a given time interval, one can estimate an evolutionary rate. These rates can be expressed as Darwins (defined as the log-difference in trait values divided by time in years) or Haldanes (defined as the difference in trait values scaled by their standard deviations divided by time in generations). Both types of rates have been calculated from both fossil data and contemporary time-series data on evolution from both islands and lab experiments. Such rates best capture evolutionary trends, where the mean value of a trait is changing in a consistent way through time (for more information see review in Harmon 2014). Rates estimated by Brownian motion are a different type of "rate", and some care must be taken to compare the two (see, e.g., Gingerich 1983). At the end of this chapter I will discuss the relationship between evolutionary rates calculated in Darwins and Haldanes with rates calculated by fitting Brownian motion models. *back to main text*

2: Note that there are more complicated optimization algorithms that are useful for more difficult problems in comparative methods. In the case presented here, where the surface is smooth and has a single peak, almost any algorithm will work. *back to main text*

References

Burnham, K. P., and D. R. Anderson. 2003. Model selection and multimodel inference: A practical Information-Theoretic approach. Springer Science & Business Media.

Felsenstein, J. 1985. Phylogenies and the comparative method. Am. Nat. 125:1–15.

FitzJohn, R. G. 2012. Diversitree: Comparative phylogenetic analyses of diversification in R. Methods Ecol. Evol. 3:1084–1092.

Freckleton, R. P. 2012. Fast likelihood calculations for comparative analyses. Methods Ecol. Evol. 3:940–947.

Garland, T., Jr. 1992. Rate tests for phenotypic evolution using phylogenetically independent contrasts. Am. Nat. 140:509–519.

Gingerich, P. D. 1983. Rates of evolution: Effects of time and temporal scaling. Science 222:159–161.

Harmon, L. J. 2014. Macroevolutionary rates. *in* The princeton guide to evolution. Princeton University Press.

Ho, L. S. T., and C. Ané. 2014. A linear-time algorithm for gaussian and non-gaussian trait evolution models. Syst. Biol. 63:397–408.

Chapter 5: Fitting Brownian Motion Models to Multiple Characters

Section 5.1: Introduction

As discussed in Chapter 4, body size is one of the most important traits of an animal. In particular, scientists often argue that body size is important because of its close relationships to almost all of an animal's ecological interactions, from whether it is a predator or prey to its metabolic rate. If that is true, we should be able to use body size to predict other traits that might be related through shared evolutionary processes. We need to understand how the evolution of body size is correlated with other species' characteristics.In this chapter, we will use the example of home range size, which is the area where an animal carries out its day-to-day activities. We will again use data from Garland (1992) and test for a relationship between body size and the size of a mammal's home range.

A wide variety of hypotheses can be framed as tests of correlations between continuously varying traits across species. For example, is the body size of a species related to its metabolic rate? How does the head length of a species relate to overall size, and do deviations from this relationship relate to an animal's diet? These questions and others like them are of interest to evolutionary biologists because they allow us to test hypotheses about the factors in influencing character evolution over long time scales. These types of approaches allow us to answer some of the classic "why" questions in biology. Why are elephants so large? Why do some species of crocodilians have longer heads than others? If we find a correlation between two characters, we might suspect that there is a causal relationship between our two variables of interest - or perhaps that both of our measured variables share a common cause.

In this chapter, I describe methods for using empirical data to estimate the parameters of multivariate Brownian motion models. I will then describe a model-fitting approach to test for evolutionary correlations. This model fitting approach is simple but not commonly used. Finally, I will review two common statistical approaches to test for evolutionary correlations, phylogenetic independent contrasts and phylogenetic generalized least squares, and describe their relationship to model-fitting approaches.

Section 5.2: What is evolutionary correlation?

There is sometimes a bit of confusion among beginners as to what, exactly, we are doing when we carry out a comparative method, especially when testing for character correlations. Common language that comparative methods "control for phylogeny" or "remove the phylogeny from the data" is not necessarily enlightening. Another common explanation is that species are not statistically independent and that we must account for that with comparative methods,

is accurate, I still don't think this statement fully captures the tree-thinking perspective enabled by comparative methods. In this section, I will use the particular example of correlated evolution to try to illustrate the power of comparative methods and how they differ from standard statistical approaches that do not use phylogenies.

In statistics, two variables can be correlated with one another. We might refer to this as a standard correlation. When two traits are correlated, it means that given the value of one trait – say, body size in mammals – one can predict the value of another – like home range area. Correlations can be positive (large values of x are associated with large values of y) or negative (large values of x are associated with small values of y). A surprisingly wide variety of hypotheses in biology can be tested by evaluating correlations between characters.

In comparative biology, we are often interested more specifically in evolutionary correlations. Evolutionary correlations occur when two traits tend to evolve together due to processes like mutation, genetic drift, or natural selection. If there is an evolutionary correlation between two characters, it means that we can predict the magnitude and direction of changes in one character given knowledge of evolutionary changes in another. Just like standard correlations, evolutionary correlations can be positive (increases in trait x are associated with increases in y) or negative (decreases in x are associated with increases in y).

We can now contrast standard correlations, testing the relationships between trait values across a set of species, with evolutionary correlations - where evolutionary changes in two traits are related to each other. This is a key distinction, because phylogenetic relatedness alone can lead to a relationship between two variables that are not, in fact, evolving together (Figure 5.1; also see Felsenstein 1985). In such cases, standard correlations will, correctly, tell us that one can predict the value of trait y by knowing the value of trait x, at least among extant species; but we would be misled if we tried to make any evolutionary causal inference from this pattern. In the example of Figure 5.1, we can only predict x from y because the value of trait x tells us which clade the species belongs to, which, in turn, allows reasonable prediction of y. In fact, this is a classical example of a case where correlation is not causation: the two variables are only correlated with one another because both are related to phylogeny.

If we want to test hypotheses about trait evolution, we should specifically test evolutionary correlations1. If we find a relationship among the independent contrasts for two characters, for example, then we can infer that changes in each character are related to changes in the other – an inference that is much closer to most biological hypotheses about why characters might be related. In this case, then, we can think of statistical comparative methods as focused on disentangling patterns due to phylogenetic relatedness from patterns due to evolutionary correlations.

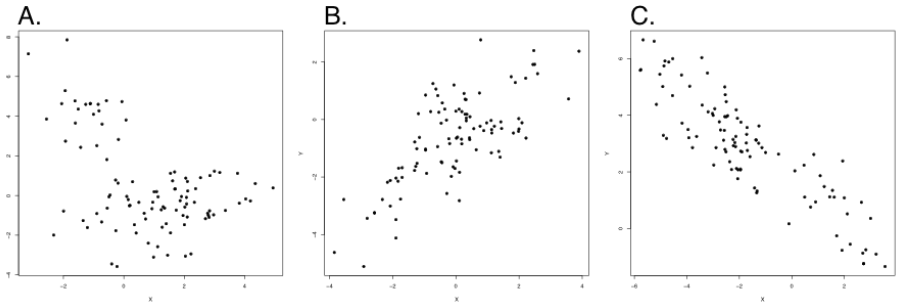

Figure 1: Figure 5.1. Examples from simulations of pure birth trees ($b = 1$) with $n = 100$ species. Plotted points represent character values for extant species in each clade. In all three panels, $\sigma_x^2 = \sigma_y^2 = 1$. σ_{xy}^2 varies with $\sigma_{xy}^2 = 0$ (panel A), $\sigma_{xy}^2 = 0.8$ (panel B), and $\sigma_{xy}^2 = -0.8$ (panel C). Note the (apparent) negative correlation in panel A, which can be explained by phylogenetic relatedness of species within two clades. Only panels B and C show data with an evolutionary correlation. However, this would be difficult or impossible to conclude without using comparative methods.

Section 5.3: Modeling the evolution of correlated characters

We can model the evolution of multiple (potentially correlated) continuous characters using a multivariate Brownian motion model. This model is similar to univariate Brownian motion (see chapter 3), but can model the evolution of many characters at the same time. As with univariate Brownian motion, trait values change randomly in both direction and distance over any time interval. Here, though, these changes are drawn from multivariate normal distributions. Multivariate Brownian motion can encompass the situation where each character evolves independently of one another, but can also describe situations where characters evolve in a correlated way.

We can describe multivariate Brownian motion with a set of parameters that are described by $\mathbf{a}$, a vector of phylogenetic means for all m characters:

(eq. 5.1)

$$\mathbf{a} = \begin{bmatrix} \bar{z}_1(0) & \bar{z}_2(0) & \dots & \bar{z}_m(0) \end{bmatrix}$$

This vector represents the starting point in m-dimensional space for our random walk. In the context of comparative methods, this is the character measurements for the lineage at the root of the tree. Additionally, we have an evolutionary rate matrix $\mathbf{R}$:

(eq. 5.2)

$$\mathbf{R} = \begin{bmatrix} \sigma_1^2 & \sigma_{21} & \dots & \sigma_{n1} \\ \sigma_{21} & \sigma_2^2 & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1n} & \dots & \dots & \sigma_m^2 \end{bmatrix}$$

Here, the rate parameter for each axis (σ_i^2) is along the matrix diagonal. Off-diagonal elements represent evolutionary covariances between pairs of axes (note that $\sigma_{ij} = \sigma_{ji}$). It is worth noting that each individual character evolves under a Brownian motion process. Covariances among characters, though, potentially make this model distinct from one where each character evolves independently of all the others (Figure 5.2).

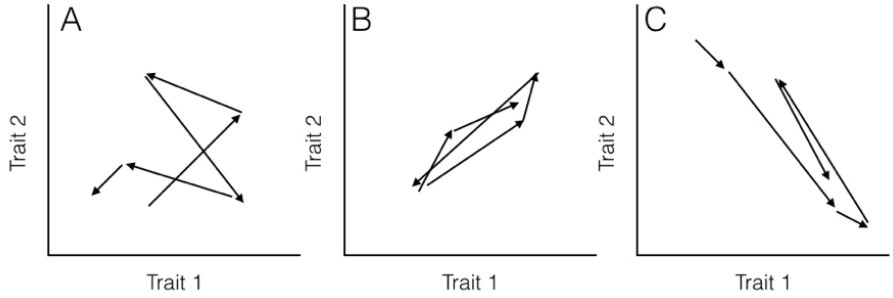

Figure 2: Figure 5.2. Hypothetical pathways of evolution (arrows) for (A) two uncorrelated traits, (B) two traits evolving with a positive covariance, and (C) two traits evolving with a negative covariance. Note that in (B), when trait 1 gets larger trait 2 also gets larger, but in (C) positive changes in trait 1 are paired with negative changes in trait 2.

When you have data for multiple continuous characters across many species along with a phylogenetic tree, you can fit a multivariate Brownian motion model to the data, as discussed in Chapter 3. The equations for estimating $\hat{\mathbf{a}}$ (the estimated vector of phylogenetic means for all characters) and $\hat{\mathbf{R}}$ (the estimated evolutionary rate matrix) are (Revell and Harmon 2008, Hohenlohe and Arnold (2008)):

(eq. 5.3)

$$\hat{\mathbf{a}} = [(\mathbf{1}\mathbf{C}^{-1}\mathbf{1})^{-1}(\mathbf{1}\mathbf{C}^{-1}\mathbf{X})]^{\intercal}$$

(eq. 5.4)

$$\hat{\mathbf{R}} = \frac{(\mathbf{X} - \mathbf{1}\hat{\mathbf{a}})^{\intercal}\mathbf{C}^{-1}(\mathbf{X} - \mathbf{1}\hat{\mathbf{a}})}{n}$$

Note here that we use $\mathbf{X}$ to denote the n (species) $\times m$ (traits) matrix of all traits across all species. Note the similarity between these multivariate equations (5.3 and 5.4) and their univariate equivalents (equations 4.6 and 4.7).

To calculate the likelihood, we can use the fact that, under our multivariate Brownian motion model, the joint distribution of all traits across all species has a multivariate normal distribution. Again following Chapter 3, we find the variance-covariance matrix that describes that model by combining the two matrices $\mathbf{R}$ and $\mathbf{C}$ into a single large matrix using the Kroeneker product:

(eq. 5.5)

$$\mathbf{V} = \mathbf{R} \otimes \mathbf{C}$$

This matrix $\mathbf{V}$ is $nm \times nm$. We can then substitute $\mathbf{V}$ for $\mathbf{C}$ in equation (4.5) to calculate the likelihood:

(eq. 5.6)

$$L(\mathbf{x}_{nm}|\mathbf{a}, \mathbf{R}, \mathbf{C}) = \frac{e^{-1/2(\mathbf{x}_{nm}-\mathbf{D}\cdot\mathbf{a})^{\intercal}(\mathbf{V})^{-1}(\mathbf{x}_{n}m-\mathbf{D}\cdot\mathbf{a})}}{\sqrt{(2\pi)^{nm}det(\mathbf{V})}}$$

Here $\mathbf{D}$ is an $nm \times m$ design matrix where each element $\mathbf{D}_{ij}$ is 1 if $(j-1)\cdot n < i \leq j \cdot n$ and 0 otherwise. is a single vector with all trait values for all species, listed so that the first n elements in the vector are trait 1, the next n are for trait 2, and so on:

(eq. 5.7)

$$\mathbf{x}_{nm} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} & x_{21} & \dots & x_{nm} \end{bmatrix}$$

Again, we can find the value of the likelihood at its maximum by calculating $L(\mathbf{x}_{nm}|\mathbf{a}, \mathbf{R}, \mathbf{C})$ using eq. 5.6.

Section 5.4: Testing for evolutionary correlations

There are many ways to test for evolutionary correlations between two characters. Traditional methods like PICs and PGLS work great for testing evolutionary regression, which is very similar to testing for evolutionary correlations. However, when using those methods the connection to actual models of character evolution can remain opaque. Thus, I will first present approaches to test for correlated evolution based on model selection using AIC and Bayesian analysis. I will then return to "standard" methods for evolutionary regression at the end of the chapter.

Section 5.4a: Testing for character correlations using maximum likelihood and AIC

To test for an evolutionary correlation between two characters, we are really interested in the elements in the matrix $\mathbf{R}$. For two characters, x and y, $\mathbf{R}$ can be written as:

(eq. 5.8)

$$\mathbf{R} = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix}$$

We are interested in the parameter σ_{xy} - the evolutionary covariance - and whether it is equal to zero (no correlation) or not. One simple way to test this hypothesis is to set up two competing hypotheses and compare them to each other. One hypothesis (H_1) is that the traits evolve independently of each other, and another (H_2) that the traits evolve with some covariance σ_{xy}. We can write these two rate matrices as:

(eq. 5.9)

$$\mathbf{R}_{H_1} = \begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{bmatrix} \quad \mathbf{R}_{H_2} = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix}$$

We can calculate an ML estimate of the parameters in $\mathbf{R}_{H_2}$ using equation 5.4. The maximum likelihood estimate of $\mathbf{R}_{H_1}$ can be obtained by noting that, if character evolution is independent across all characters, then both σ_x^2 and σ_y^2 can be obtained by treating each character separately and using equations from chapter 3 to solve for each. It turns out that the ML estimates for σ_x^2 and σ_y^2 are always exactly the same for H_1 and H_2.

To compare these two models, we calculate the likelihood of each using equation 5.5. We can then compare these two likelihoods using either a likelihood ratio test or by comparing AICc scores (see chapter 2.

For the mammal example, we can consider the two traits of (ln-transformed) body size and home range size (Garland 1992). These two characters have a positive correlation using standard regression analysis ($r = 0.27$), and a linear regression is significant ($P = 0.0001$; Figure 5.3). If we fit a multivariate Brownian motion model to these data, considering home range as trait 1 and body mass as trait 2, we obtain the following parameter estimates:

(eq. 5.10)

$$\hat{\mathbf{a}}_{H_2} = \begin{bmatrix} 2.54 \\ 4.64 \end{bmatrix} \quad \hat{\mathbf{R}}_{H_2} = \begin{bmatrix} 0.24 & 0.10 \\ 0.10 & 0.09 \end{bmatrix}$$

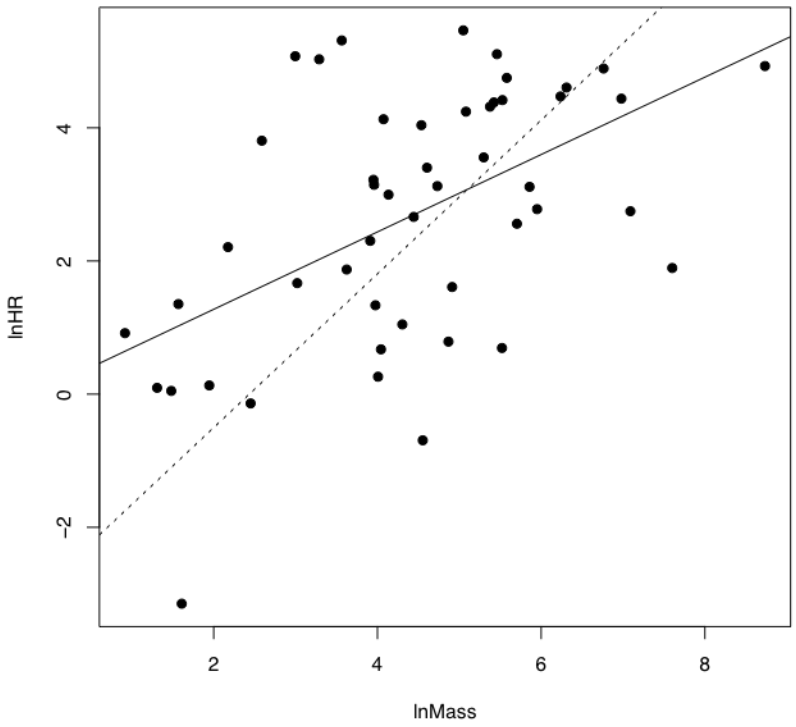

Figure 3: Figure 5.3. The relationship between mammal body mass and home-range size. Solid line is a regression line from a standard analysis, dotted line from PGLS, which uses the phylogenetic tree (see below for a detailed description).

Note the positive off-diagonal element in the estimated **R** matrix, suggesting a positive evolutionary correlation between these two traits. This model corresponds to hypothesis 2 above, and has a log-likelihood of $lnL = -164.0$. If we fit a model with no correlation between the two traits, we obtain:

(eq. 5.11)

$$\hat{\mathbf{a}}_{H_2} = \begin{bmatrix} 2.54 \\ 4.64 \end{bmatrix} \quad \hat{\mathbf{R}}_{H_2} = \begin{bmatrix} 0.24 & 0 \\ 0 & 0.09 \end{bmatrix}$$

It is worth noting again that only the estimates of the evolutionary correlation were affected by this model restriction; all other parameter estimates remain the same. This model has a more negative log-likelihood of $lnL = -180.5$.

A likelihood ratio test gives $\Delta = 33.0$, and $P << 0.001$, rejecting the null hypothesis. The difference in AIC_c scores is 30.9, and the Akaike weight for model 2 is effectively 1.0. Both ways of comparing these two models give strong support for hypothesis 2. We can conclude that there is an evolutionary correlation between body mass and home range size in mammals. What this means in evolutionary terms is that, across mammals, evolutionary changes in body mass tend to covary with changes in home range.

Section 5.4b: Testing for character correlations using Bayesian model selection

We can also implement a Bayesian approach to testing for the correlated evolution of two characters. The simplest way to do this is just to use the standard algorithm for Bayesian MCMC to fit a correlated model to the two characters. We can modify the algorithm presented in chapter 2 as follows:

1. Sample a set of starting parameter values σ_x^2, σ_y^2, and σ_{xy} from the prior distribution. For this example, we can set our prior distribution as uniform between 0 and 1 for σ_x^2 and σ_y^2 and uniform from -1 to +1 for σ_{xy}.
2. Given the current parameter values, select new proposed parameter values using the proposal density $Q(p'|p)$. Here, for all three parameter values, we will use a uniform proposal density with width 0.2, so that $Q(p'|p) \sim U(p - 0.1, p + 0.1)$.
3. Calculate three ratios:

a. The prior odds ratio. This is the ratio of the probability of drawing the parameter values p and p' from the prior. Since our priors are uniform, this is always 1.
b. The proposal density ratio. This is the ratio of probability of proposals going from p to p' and the reverse. Our proposal density is symmetrical, so that $Q(p'|p) = Q(p|p')$ and $a_2 = 1$.

c. The likelihood ratio. This is the ratio of probabilities of the data given the two different parameter values. We can calculate these probabilities from equation 5.6 above. (eq. 5.12)

$$a_3 = \frac{L(p'|D)}{L(p|D)} = \frac{P(D|p')}{P(D|p)}$$

4. Find the product of the prior odds, proposal density ratio, and the likelihood ratio. In this case, both the prior odds and proposal density ratios are 1, so $a = a_3$.
5. Draw a random number x from a uniform distribution between 0 and 1. If $x < a$, accept the proposed value of all parameters; otherwise reject, and retain the current parameter values.
6. Repeat steps 2-5 a large number of times.

We can then inspect the posterior distribution for the parameter is significantly greater than (or less than) zero. As an example, I ran this MCMC for 100,000 generations, discarding the first 10,000 generations as burn-in. I then sampled the posterior distribution every 100 generations, and obtained the following parameter estimates: $\sigma_x^2 = 0.26$ (95% CI: 0.18 - 0.38), $\sigma_y^2 = 0.10$ (95% CI: 0.06 -0.15), and $\sigma_{xy} = 0.11$ (95% CI: 0.06 - 0.17; see Figure 5.4). These results are comparable to our ML estimates. Furthermore, the 95% CI for σ_{xy} does not overlap with 0; in fact, none of the 901 posterior estimates of σ_{xy} are less than zero. Again, we can conclude with confidence that there is an evolutionary correlation between these two characters.

Section 5.5c: Testing for character correlations using traditional approaches (PIC, PGLS)

The approach outlined above, which tests for an evolutionary correlation among characters using model selection, is not typically applied in the comparative biology literature. Instead, most tests of character correlation rely on phylogenetic regression using one of two methods: phylogenetic independent contrasts (PICs) and phylogenetic general least squares (PGLS). PGLS is actually mathematically identical to PICs in the simple case described here, and more flexible than PICs for other models and types of characters. Here I will review both PICs and PGLS and explain how they work and how they relate to the models described above.

Phylogenetic independent contrasts can be used to carry out a regression test for the relationship between two different characters. To do this, one calculates standardized PICs for trait x and trait y. One then uses standard regression forced through the origin to test for a relationship between these two sets of PICs. It is necessary to force the regression through the origin because the direction of subtraction of contrasts across any node in the tree is arbitrary; a reflection of all of the contrasts across both axes simultaneously should have no effect on the analyses2.

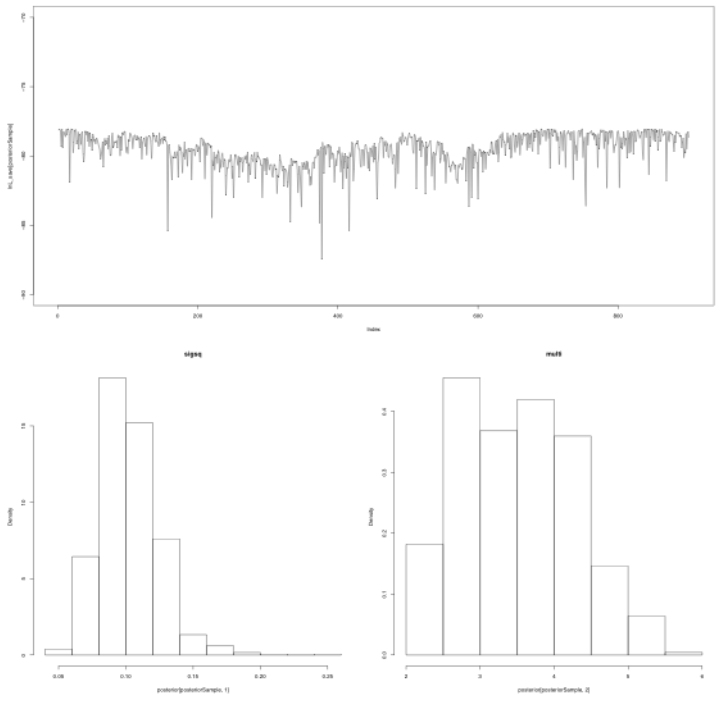

Figure 4: Figure 5.4. Bayesian analysis of evolutionary correlation. A. likelihood trace, B. posterior distribution of σ_{xy}, C. posterior distribution of a_2.

For mammal homerange and body mass, a PIC regression test shows a significant correlation between the two traits ($P << 0.0001$; Figure 5.5).

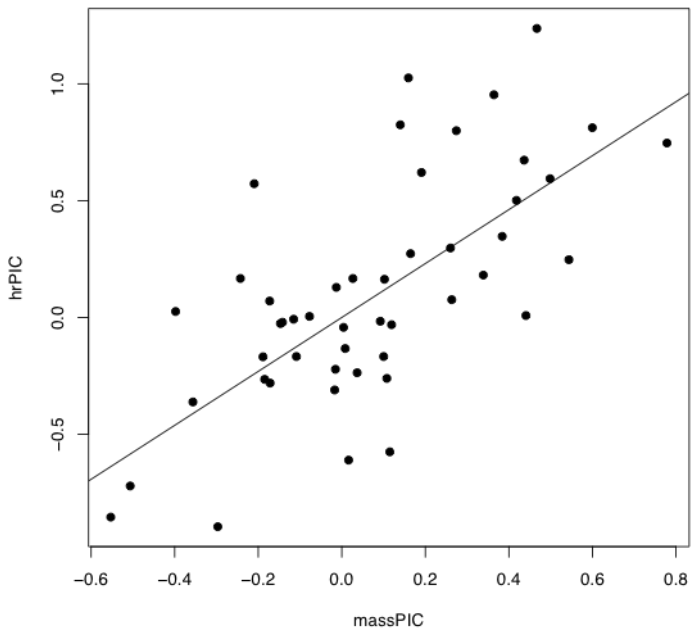

Figure 5: Figure 5.5. Regression based on independent contrasts. The regression line is forced through the origin.

There is one drawback to PIC regression analysis, though – one does not recover an estimate of the intercept of the regression of y on x – that is, the value of y one would expect when $x = 0$. The easiest way to get this parameter estimate is to instead use Phylogenetic Generalized Least Squares (PGLS). PGLS uses the common statistical machinery of generalized least squares, and applies it to phylogenetic comparative data. In normal generalized least squares, one constructs a model of the relationship between y and x, as:

(eq. 5.13)

$$\mathbf{y} = \mathbf{X_D b} + \epsilon$$

Here, $\mathbf{y}$ is an $n \times 1$ vector of trait values and $\mathbf{b}$ is a vector of unknown regression coefficients that must be estimated from the data. $\mathbf{X_D}$ is a design matrix including the traits that one wishes to test for a correlation with y and – if the model includes an intercept – a column of 1s. To test for correlations, we use:

(eq. 5.14)

$$\mathbf{X_D} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{bmatrix}$$

In this case, b is 2×1 and the resulting model can be used to test correlations between two characters. However, $\mathbf{X_D}$ could also be multivariate, and can include more than one character that might be related to y. This allows us to carry out the equivalent of multiple regression in a phylogenetic context. Finally, ϵ are the residuals – the difference between the y-values predicted by the model and their actual values. In traditional regression, one assumes that the residuals are all normally distributed with the same variance. By contrast, with GLS, one assumes that the residuals might not be independent of each other; instead, they are multivariate normal with expected mean zero and some variance-covariance matrix $\mathbf{\Omega}$.

In the case of Brownian motion, we can model the residuals as having variances and covariances that follow the structure of the phylogenetic tree. In other words, we can substitute our phylogenetic variance-covariance matrix $\mathbf{C}$ as the matrix $\mathbf{\Omega}$. We can then carry out standard GLS analyses to estimate model parameters:

(eq. 5.15)

$$\hat{\mathbf{b}} = (\mathbf{X}_D^\intercal \mathbf{\Omega}^{-1} \mathbf{X}_D^\intercal)^{-1} \mathbf{X}_D^\intercal \mathbf{\Omega}^{-1} \mathbf{y} = (\mathbf{X}_D^\intercal \mathbf{C}^{-1} \mathbf{X}_D^\intercal)^{-1} \mathbf{X}_D^\intercal \mathbf{C}^{-1} \mathbf{y}$$

One might notice a similarity between equation 5.15 and equation 4.7. In fact, if $\mathbf{X}_D$ from (5.14) is used for PGLS, then the first term in $\hat{\mathbf{b}}$ is the phylogenetic mean θ. The other term in $\hat{\mathbf{b}}$ will be an estimate for the slope of the relationship between y and x, the calculation of which statistically controls for the effect of phylogenetic relationships.

Applying PGLS to mammal body mass and home range results in an identical estimate of the slope and P-value as we obtain using independent contrasts (see Box 4.1). PGLS also returns an estimate of the intercept of this relationship, which cannot be obtained from the PICs.

Of course, another difference is that PICs and PGLS use regression, while the approach outlined above tests for a correlation. These two types of statistical tests are different. Correlation tests for a relationship between x and y, while regression tries to find the best way to predict y from x. For correlation, it does not matter which variable we call x and which we call y. However, in regression we will get a different slope if we predict y given x instead of predicting x given y. The model that is assumed by phylogenetic regression models is also different from the model above, where we assumed that the two characters evolve under a

correlated Brownian motion model. By contrast, PGLS (and, implicitly, PICs) assume that the deviations of each species from the regression line evolve under a Brownian motion model. We can imagine, for example, that species can freely slide along the regression line, but that evolving around that line can be captured by a normal Brownian model. Another way to think about a PGLS model is that we are treating x as a fixed property of species. The deviation of y from what is predicted by x is what evolves under a Brownian motion model. If this seems strange, that's because it is! There are other, more complex models for modeling the correlated evolution of two characters that make assumptions that are more evolutionarily realistic; we will return to this topic later in the book. At the same time, PGLS is a well-used method for evolutionary regression, and is undoubtedly useful despite its somewhat strange assumptions.

PGLS analysis, as described above, assumes that characters are evolving under a Brownian motion model. However, one can change the structure of the error variance-covariance matrix to reflect other models of evolution, such as OU. We return to this topic in a later chapter.

Section 5.6: Summary

There are at least four methods for testing for an evolutionary correlation between continuous characters: likelihood ratio test, AIC model selection, PICs, and PGLS. These four methods as presented all make the same assumptions about the data and, therefore, have quite similar statistical properties (even simulating under a multivariate Brownian motion model, which deviates from the model assumptions, both PICs and PGLS have appropriate Type I error rates and very similar power). Any of these are good choices for testing for the presence of an evolutionary correlation in your data.

Section 5.7: Footnotes

1: We might also want to carry out linear regression, which is related to correlation analysis but distinct. We will show examples of phylogenetic regression at the end of this chapter.*back to main text*

2: Another way to think about regression through the origin is to think of pairs of contrasts across any node in the tree as two-dimensional vectors. Calculating a vector correlation is equivalent to calculating a regression forced through the origin. *back to main text*

Section 5.8: References

Felsenstein, J. 1985. Phylogenies and the comparative method. Am. Nat. 125:1–15.

Garland, T., Jr. 1992. Rate tests for phenotypic evolution using phylogenetically independent contrasts. Am. Nat. 140:509–519.

Hohenlohe, P. A., and S. J. Arnold. 2008. MIPoD: A hypothesis-testing framework for microevolutionary inference from patterns of divergence. Am. Nat. 171:366–385.

Revell, L. J., and L. J. Harmon. 2008. Testing quantitative genetic hypotheses about the evolutionary rate matrix for continuous characters. Evol. Ecol. Res. 10:311–331.

Chapter 6: Beyond Brownian Motion

Section 6.1: Introduction

Detailed studies of contemporary evolution have revealed a rich variety of processes that influence how traits evolve through time. Consider the famous studies of Darwin's finches, *Geospiza*, in the Galapagos islands carried out by Peter and Rosemary Grant, among others (e.g. Grant and Rosemary Grant 2011). These studies have documented the action of natural selection on traits from one generation to the next. One can see very clearly how changes in climate – especially the amount of rainfall – affect the availability of different types of seeds (Grant and Grant 2002). These changing resources in turn affect which individuals survive within the population. When natural selection acts on traits that can be inherited from parents to offspring, those traits evolve.

One can obtain a dataset of morphological traits, including measurements of body and beak size and shape, along with a phylogenetic tree for several species of Darwin's finches. Imagine that you have the goal of analyzing the tempo and mode of morphological evolution across these species of finch. We can start by fitting a Brownian motion model to these data. However, a Brownian model (which, as we learned in Chapter 3, corresponds to a few simple scenarios of trait evolution) hardly seems realistic for a group of finches known to be under strong and predictable directional selection.

Brownian motion is very commonly in comparative biology: in fact, a large number of comparative methods that researchers use for continuous traits assumes that those traits evolve under a Brownian motion model. The scope of other models beyond Brownian motion that we can use to model continuous trait data on trees is somewhat limited. However, some methods have been developed that break free of this limitation, moving the field beyond Brownian motion. In this chapter I will discuss these new approaches and what they can tell us about evolution. I will also describe how moving beyond Brownian motion can point the way forward for statistical comparative methods.

In this chapter, I will consider four ways that comparative methods can move beyond simple Brownian motion models: by transforming the variance-covariance matrix describing trait covariation among species, by incorporating variation in rates of evolution, by accounting for evolutionary constraints, and by modeling adaptive radiation, species interactions, and other biological processes. It should be apparent that the models listed here do not span the complete range of possibilities, and so my list is not meant to be comprehensive. Instead, I hope that readers will view these as examples, and that future researchers will add to this list and enrich the set of models that we can fit to data.

Section 6.2: Transforming the evolutionary variance-covariance matrix

In 1999, Mark Pagel introduced three statistical models that allow one to test whether data deviates from a constant-rate Mk process evolving on a phylogenetic tree (Pagel 1999a, Pagel (1999b)). Each of these three models is a statistical transformation of the elements of the phylogenetic variance-covariance matrix, **C**, that we first encountered in Chapter 3. All three can also be thought of as a transformation of the branch lengths of the tree, which adds a more intuitive understanding of the statistical properties of the tree transformations (Figure 6.1). We can transform the tree and then simulate characters under a Brownian motion model on the transformed tree, generating very different patterns than if they had been simulated on the starting tree.

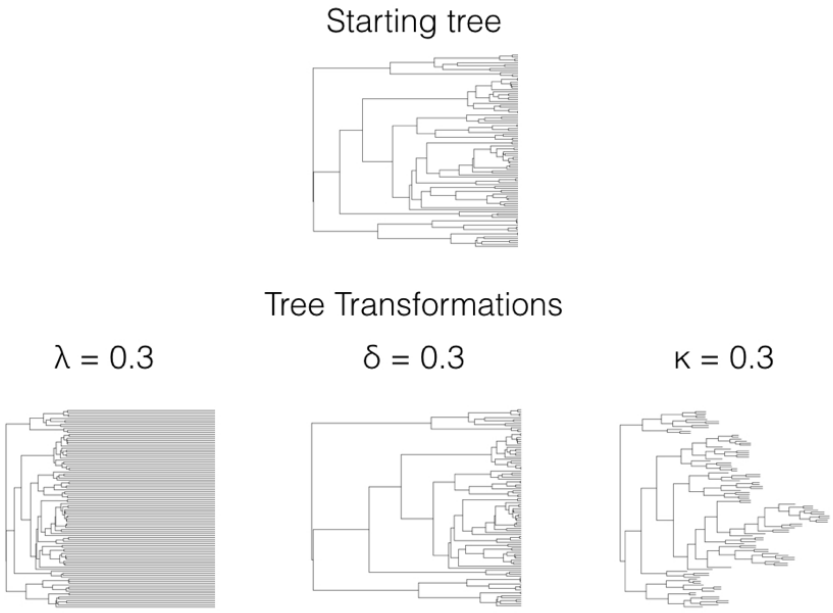

Figure 1: Figure 6.1. Branch length transformations effectively alter the relative rate of evolution on certain branches in the tree. If we make a branch longer, there is more "evolutionary time" for characters to change, and so we are effectively increasing the rate of evolution along that branch.

There are three Pagel tree transformations (lambda: λ, delta: δ, and kappa: κ). I will describe each of them along with common methods for fitting Pagel models under ML, AIC, and Bayesian frameworks. Pagel's three transformations can also be related to evolutionary processes, although those relationships

are sometimes vague compared to approaches based on explicit evolutionary models rather than tree transformations (see below for more comments on this distinction).

Perhaps the most commonly used Pagel tree transformation is λ. When using λ, one multiplies all off-diagonal elements in the phylogenetic variance-covariance matrix by the value of λ. The diagonal elements remain unchanged. So, if the original matrix is:

(Equation 6.1)

$$\mathbf{C_o} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_n^2 \end{bmatrix}$$

Then the transformed matrix will be:

(Equation 6.2)

$$\mathbf{C}_\lambda = \begin{bmatrix} \sigma_1^2 & \lambda \cdot \sigma_{12} & \dots & \lambda \cdot \sigma_{1n} \\ \lambda \cdot \sigma_{21} & \sigma_2^2 & \dots & \lambda \cdot \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda \cdot \sigma_{n1} & \lambda \cdot \sigma_{n2} & \dots & \sigma_n^2 \end{bmatrix}$$

In terms of branch length transformations, λ compresses internal branches while leaving the tip branches of the tree unaffected (Figure 6.1). can range from 1 (no transformation) to 0 (which results in a complete star phylogeny, with all tip branches equal in length and all internal branches of length 0). One can use values of λ greater than one on the variance-covariance matrix, although some values of λ result in matrices that are not valid variance-covariance matrices and/or do not correspond with any phylogenetic tree transformation. For this reason I recommend that λ be limited to values between 0 and 1.

λ is often used to measure the "phylogenetic signal" in comparative data. This makes intuitive sense, as λ scales the tree between a constant-rates model to one where every species is statistically independent of every other species in the tree. Statistically, this can be very useful information. However, there is some danger is in attributing a statistical result – either phylogenetic signal or not – to any particular biological process. For example, phylogenetic signal is sometimes called a "phylogenetic constraint." But one way to obtain a high phylogenetic signal (λ near 1) is to evolve traits under a Brownian motion model, which involves completely unconstrained character evolution. Likewise, a lack of phylogenetic signal – which might be called "low phylogenetic constraint" – results from an OU model with a high α parameter (see below), which is a model where trait evolution away from the optimal value is, in fact, highly

constrained. Revell et al. (2008) show a broad range of circumstances that can lead to patterns of high or low phylogenetic signal, and caution against over-interpretation of results from analyses of phylogenetic signal, like Pagel's λ. Also worth noting is that statistical estimates of λ under a ML model tend to be clustered near 0 and 1 regardless of the true value, and AIC model selection can tend to prefer λ models even when data is simulated under Brownian motion (Boettiger et al. 2012).

Pagel's δ is designed to capture variation in rates of evolution through time. Under the delta transformation, all elements of the phylogenetic variance-covariance matrix are raised to the power δ. So, if our original C matrix is given above (equation 6.1), then the -transformed version will be:

(6.3)

$$\mathbf{C}_\delta = \begin{bmatrix} (\sigma_1^2)^\delta & (\sigma_{12})^\delta & \dots & (\sigma_{1n})^\delta \\ (\sigma_{21})^\delta & (\sigma_2^2)^\delta & \dots & (\sigma_{2n})^\delta \\ \vdots & \vdots & \ddots & \vdots \\ (\sigma_{n1})^\delta & (\sigma_{n2})^\delta & \dots & (\sigma_n^2)^\delta \end{bmatrix}$$

Since these elements represent the heights of nodes in the phylogenetic tree, then δ can also be viewed as a transformation of phylogenetic node heights. When δ is one, the tree is unchanged and one still has a constant-rate Brownian motion process; when δ is less than 1, node heights are reduced, but deeper branches in the tree are reduced less than shallower branches (Figure 6.1). This effectively represents a model where the rate of evolution slows through time. By contrast, $\delta > 1$ stretches the shallower branches in the tree more than the deep branches, mimicking a model where the rate of evolution speeds up through time. There is a close connection between the delta model, the ACDC model (Blomberg et al. 2003), and Harmon et al.'s (2010) early burst model [see also Uyeda and Harmon (2014), especially the appendix).

Finally, the κ transformation is sometimes used to capture patterns of "speciational" change in trees. In the κ model, one raises all of the branch lengths in the tree by the power κ. This has a complicated effect on the phylogenetic variance-covariance matrix, as the effect that this transformation has on each covariance element depends on both the value of κ and the number of branches that extend from the root of the tree to the most recent common ancestor of each pair of species. So, if our original C matrix is given by equation 6.1, the transformed version will be:

(Equation 6.4)

$$\mathbf{C_o} = \begin{pmatrix} b^k_{1,1}+b^k_{1,2}\cdots+b^k_{1,d_1} & b^k_{1-2,1}+b^k_{1-2,2}\cdots+b^k_{1-2,d_{1-2}} & \cdots & b^k_{1-n,1}+b^k_{1-n,2}\cdots+b^k_{1-n,d_{1-n}} \\ b^k_{2-1,1}+b^k_{2-1,2}\cdots+b^k_{2-1,d_{1-2}} & b^k_{2,1}+b^k_{2,2}\cdots+b^k_{2,d_2} & \cdots & b^k_{2-n,1}+b^k_{2-n,2}\cdots+b^k_{2-n,d_{2-n}} \\ \vdots & \vdots & \ddots & \vdots \\ b^k_{n-1,1}+b^k_{n-1,2}\cdots+b^k_{n-1,d_{1-n}} & b^k_{n-2,1}+b^k_{n-2,2}\cdots+b^k_{n-2,d_{1-2}} & \cdots & b^k_{n,1}+b^k_{n,2}\cdots+b^k_{n,d_n} \end{pmatrix}$$

where $b_{x,y}$ is the branch length of the branch that is the most recent common ancestor of taxa x and y, while $d_{x,y}$ is the total number of branches that one encounters traversing the path from the root to the most recent common ancestor of the species pair specified by x, y (or to the tip x if just one taxon is specified). Needless to say, this transformation is easier to understand as a transformation of the tree branches themselves rather than of the associated variance-covariance matrix.

When the κ parameter is one, the tree is unchanged and one still has a constant-rate Brownian motion process; when $\kappa = 0$, all branch lengths are one. Kappa values in between these two extremes represent intermediates (Figure 6.1). Kappa is often interpreted in terms of a model where character change is more or less concentrated at speciation events. For this interpretation to be valid, we have to assume that the phylogenetic tree, as given, includes all (or even most) of the speciation events in the history of the clade. The problem with this assumption is that speciation events are almost certainly missing due to sampling: perhaps some living species from the clade have not been sampled, or species that are part of the clade have gone extinct before the present day and are thus not sampled. There are much better ways of estimating speciational models that can account for these issues in sampling (e.g. Bokma 2008, Goldberg and Igić (2012)); these newer methods should be preferred over Pagel's κ for testing for a speciational pattern in trait data.

There are two main ways to assess the fit of the three Pagel-style models to data. First, one can use ML to estimate parameters and likelihood ratio tests (or AIC_c scores) to compare the fit of various models. As mentioned above, simulation studies suggest that this can sometimes lead to overconfidence, at least for the λ model. Sometimes researchers will compare the fit of a particular model (e.g. λ) with models where that parameter is fixed at its two extreme values (0 or 1; this is not possible with δ). Second, one can use Bayesian methods to estimate posterior distributions of parameter values, then inspect those distributions to see if they overlap with values of interest (say, 0 or 1).

We can apply these three Pagel models to the mammal body size data discussed in chapter 5, comparing the AIC_c scores for Brownian motion to that from the three transformations. We obtain the following results:

Model	Parameter estimates	ln-Likelihood	AIC_c
Brownian motion	$\sigma^2 = 0.088$, $\theta = 4.64$	-78.0	160.4
lambda	$\sigma^2 = 0.085$, $\theta = 4.64$, $\lambda = 1.0$	-78.0	162.6

Model	Parameter estimates	ln-Likelihood	AIC_c
delta	$\sigma^2 = 0.063$, $\theta = 4.60$, $\delta = 1.5$	-77.7	162.0
kappa	$\sigma^2 = 0.170$, $\theta = 4.64$, $\kappa = 0.66$	-77.3	161.1

Note that Brownian motion is the preferred model with the lowest AIC_c score, but also that all four AIC_c scores are within 3 units – meaning that we cannot easily distinguish among them as models for our mammal data.

Section 6.3: Variation in rates of trait evolution across clades

One assumption of Brownian motion is that the rate of change (σ^2) is constant, both through time and across lineages. However, some of the most interesting hypotheses in evolution relate to differences in the rates of character change across clades. For example, key innovations are evolutionary events that open up new areas of niche space to evolving clades (reviewed in Alfaro 2013, Hunter (1998)). This new niche space is an ecological opportunity that can then be filled by newly evolved species (Yoder et al. 2010). If this were happening in a clade, we might expect that rates of trait evolution would be elevated following the acquisition of the key innovation (Yoder et al. 2010).

There are several methods that one can use to test for differences in the rate of evolution across clades. First, one can compare the magnitude of independent contrasts across clades; second, one can use model comparison approaches to compare the fit of single- and multiple-rate models to data on trees; and third, one can use a Bayesian approach combined with reversible-jump machinery to try to find the places on the tree where rate shifts have occurred. I will explain each of these methods in turn.

Section 6.3a: Rate tests using phylogenetic independent contrasts

One of the earliest methods for comparing rates across clades is to compare the magnitude of independent contrasts calculated in each clade (e.g. Garland 1992). To do this, one first calculates standardized independent contrasts, separating those contrasts that are calculated within each clade of interest. As we noted in Chapter 5, these contrasts have arbitrary sign (positive or negative) but if they are squared, represent independent estimates of the Brownian motion rate parameter (σ^2). Therefore, one can compare the magnitude of independent contrasts within the clade of interest to the contrasts in another clade (or in the rest of the tree) as a test for differences in the rate of evolution (Garland 1992).

In his original description of this approach, Garland (1992) proposed using a statistical test to compare the absolute value of contrasts between clades (or

between a single clade and the rest of the phylogenetic tree). In particular, Garland (1992) suggests using a t-test, as long as the absolute value of independent contrasts are approximately normally distributed. However, under a Brownian motion model, the contrasts themselves – but not the absolute values of the contrasts – should be approximately normal, so it is quite likely that absolute values of contrasts will strongly violate the assumptions of a t-test.

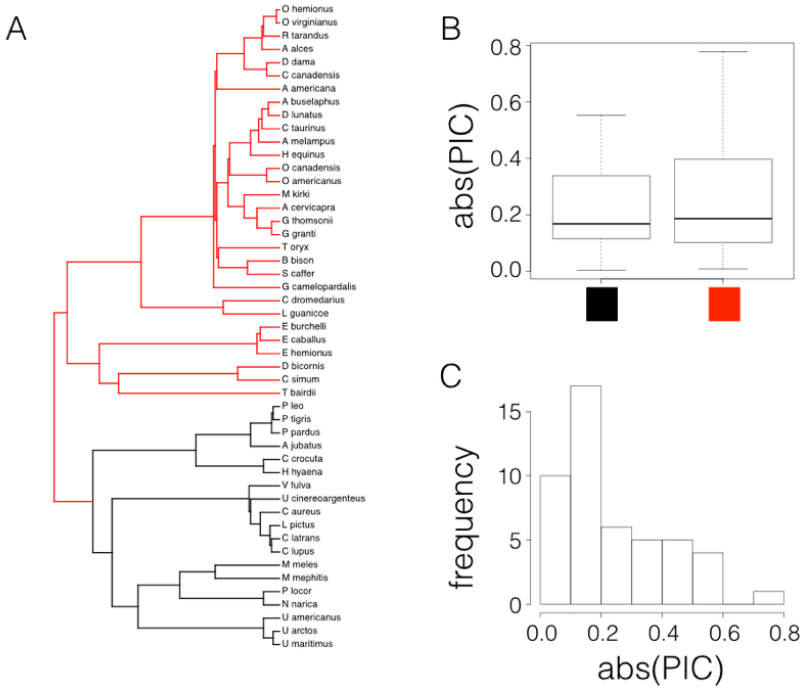

Figure 2: Figure 6.2. Rate tests comparing carnivores (black) with other mammals (red; Panel A). Box-plots show only a slight difference in the absolute value of independent contrasts for the two clades, and the distribution of absolute values of contrasts is strongly skewed.

In fact, if we try this test on mammal body size, contrasting the two major clades in the tree (carnivores versus non-carnivores, Figure 6.2A), there looks to be a small difference in the absolute value of contrasts (Figure 6.2B). A t-test is not significant (Welch two-sample t-test $P = 0.42$), but we also can see that the distribution of PIC absolute values is strongly skewed (Figure 6.2C).

There are other simple options that might work better in general. For example, one could also compare the magnitudes of the squared contrasts, although these are also not expected to follow a normal distribution. Alternatively, we can again follow Garland's (1992) suggestion and use a Mann-Whitney U-test, the nonparametric equivalent of a t-test, on the absolute values of the contrasts. Since Mann-Whitney U tests use ranks instead of values, this approach will not

be sensitive to the fact that the absolute values of contrasts are not normal. If the P-value is significant for this test then we have evidence that the rate of evolution is greater in one part of the tree than another.

In the case of mammals, a Mann-Whitney U test also shows no significant differences in rates of evolution between carnivores and other mammals ($W = 251$, $P = 0.70$).

Section 6.3b: Rate tests using maximum likelihood and AIC

One can also carry out rate comparisons using a model-selection framework (O'Meara et al. 2006, Thomas et al. (2006)). To do this, we can fit single- and multiple-rate Brownian motion models to a phylogenetic tree, then compare them using a model selection method like AIC_c. For example, in the example above, we tested whether or not one subclade in the mammal tree (carnivores) has a very different rate of body size evolution than the rest of the clade. We can use an ML-based model selection method to compare the fit of a single-rate model to a model where the evolutionary rate in carnivores is different from the rest of the clade, and use this test evaluate the support for that hypothesis.

This test requires the likelihood for a multi-rate Brownian motion model on a phylogenetic tree. We can derive such an equation using equations that are closely related to the likelihood equations presented in Chapter 4. Recall that the likelihood equations for (constant-rate) Brownian motion use a phylogenetic variance-covariance matrix, $\mathbf{C}$, that is based on the branch lengths and topology of the tree. For single-rate Brownian motion, the elements in $\mathbf{C}$ are derived from the branch lengths in the tree. Traits are drawn from a multivariate normal distribution with variance-covariance matrix:

(6.5)

$$\mathbf{V}_{H_1} = \sigma^2 \mathbf{C}_{tree}$$

One simple way to fit a multi-rate Brownian motion model is to construct separate $\mathbf{C}$ matrices, one for each rate category in the tree. For example, imagine that most of a clade evolves under a Brownian motion model with rate σ_1^2, but one clade in the tree evolves at a different (higher or lower) rate, σ_2^2. One can construct two $\mathbf{C}$ matrices: the first matrix, $\mathbf{C}_1$, includes branches that evolve under rate σ_1^2, while the second, $\mathbf{C}_2$, includes only branches that evolve under rate σ_2^2. Since all branches in the tree are included in one of these two categories, it will be true that $\mathbf{C}_{tree} = \mathbf{C}_1 + \mathbf{C}_2$. For any particular values of these two rates, traits are drawn from a multivariate normal distribution with variance-covariance matrix:

(6.6)

$$\mathbf{V}_{H_2} = \sigma_1^2\mathbf{C}_1 + \sigma_2^2\mathbf{C}_2$$

We can now treat this as a model comparison-problem, contrasting H_1: traits on the tree evolved under a constant-rate Brownian motion model, with H_2: traits on the tree evolved under a multi-rate Brownian motion model. Note that H_1 is a special case of H_2 when $\sigma_1^2 = \sigma_2^2$; that is, these two models are nested and can be compared using a likelihood ratio test. Of course, one can also compare the two models using AIC.

For the mammal body size example, you might recall our ML single-rate Brownian motion model ($\sigma^2 = 0.088$, $\bar{z}(0) = 4.64$, $lnL = -78.0$, $AIC_c = 160.4$). We can compare that to the fit of a model where carnivores get their own rate parameter (σ_c^2) that might differ from that of the rest of the tree (σ_o^2). Fitting that model, we find the following maximum likelihood parameter estimates: $\sigma_c^2 = 0.068$, $\sigma_o^2 = 0.01$, $\bar{z}(0) = 4.51$). Carnivores do appear to be evolving more rapidly. However, the fit of this model is not substantially better than the single-rate Brownian motion ($lnL = -77.6$, $AIC_c = 162.3$).

There is one complication, which is how to deal with the actual branch along which the rate shift is thought to have occurred. O'Meara et al. (2006) describe "censored" and "noncensored" versions of their test, which differ in whether or not branches where rate shifts actually occur are included in the calculation. In the censored version of the test, we omit the branch where we think a shift occurred, while in the noncensored version we include that branch in one of the two rate categories (this is what I did in the example above, adding the stem branch of carnivores in the "non-carnivore" category). One could also specify where, exactly, the rate shift occurred along the branch in question, placing part of the branch in each of the two rate categories as appropriate. However, since we typically have little information about what happened on particular branches in a phylogenetic tree, results from these two approaches are not very different – unless, as stated by O'Meara et al. (2006), unusual evolutionary processes have occurred on the branch in question.

A similar approach was described by Thomas et al. (2006) but considers differences across clades to include changes in any of the two parameters of a Brownian motion model (σ^2, $\bar{z}(0)$, or both). Remember that $\bar{z}(0)$ is the expected mean of species within a clade under a Brownian motion, but also represents the starting value of the trait at time zero. Allowing $\bar{z}(0)$ to vary across clades effectively allows different clades to have different "starting points" in phenotype space. In the case of comparing a monophyletic subclade to the rest of a tree, Thomas et al.'s (2006) approach is equivalent to the "censored" test described above. However, one drawback to both the Thomas et al. (2006) approach and the "censored" test is that, because clades each have their own mean, we no longer can tie the model that we fit using likelihood to any particular evolutionary process. Mathematically, changing $\bar{z}(0)$ in a subclade postulates that the trait value changed somehow along the branch leading to that clade, but we do not

specify the way that the trait changed – the change could have been gradual or instantaneous, and no amount or pattern of change is more or less likely than anything else. Of course, one can describe evolutionary scenarios that might act like this process - but we begin to lose any potential tie to quantitative genetic processes.

Section 6.3c: Rate tests using Bayesian MCMC

It is also possible to carry out this test in a Bayesian MCMC framework. The simplest way to do that would be to fit model H2 above, that traits on the tree evolved under a multi-rate Brownian motion model, in a Bayesian framework. We can then specify prior distributions and sample the three model parameters ($\bar{z}(0)$, σ_1^2, and σ_2^2) through our MCMC. At the end of our analysis, we will have posterior distributions for the three model parameters. We can test whether rates differ among clades by calculating a posterior distribution for the composite parameter $\sigma_{diff}^2 = \sigma_1^2 - \sigma_2^2$. The proportion of the posterior distribution for σ_{diff}^2 that is positive or negative gives the posterior probability that σ_1^2 is greater or less than σ_1^2, respectively.

Perhaps, though, researchers are unsure of where, exactly, the rate shift might have occurred, and want to incorporate some uncertainty in their analysis. In some cases, rate shifts are thought to be associated with some other discrete character, such as living on land (state 0) or in the water (1). In such cases, one way to proceed is to use stochastic character mapping (see Chapter 9) to map state changes for the discrete character on the tree, and then run an analysis where rates of evolution of the continuous character of interest depend on the mapping of our discrete states. This protocol is described most fully by Revell (2013), who also points out that rate estimates are biased to be more similar when the discrete character evolves quickly.

It is even possible to explore variation in Brownian rates without invoking particular a priori hypotheses about where the rates might change along branches in a tree. These methods rely on reversible-jump MCMC, a Bayesian statistical technique that allows one to consider a large number of models, all with different numbers of parameters, in a single Bayesian analysis. In this case, we consider models where each branch in the tree can potentially have its own Brownian rate parameter. By constraining sets of these rate parameters to be equal to one another, we can specify a huge number of models for rate variation across trees. The reversible-jump machinery, which is beyond the scope of this book, allows us to generate a posterior distribution that spans this large set of models (see Eastman et al. 2011 for details).

Section 6.4: Evolution under stabilizing selection

We can also consider the case where a trait evolves under the influence of stabilizing selection. Assume that a trait has some optimal value, and that when the population mean differs from the optimum the population will experience selection towards the optimum. As I will show below, when traits evolve under stabilizing selection with a constant optimum, the pattern of traits through time can be described under an OU model. It is worth mentioning, though, that this is only one (of many!) models that follow an OU process over long time scales. In other words, even though this model can be described by OU, we cannot make inferences the other direction and claim that OU means that our population is under constant stabilizing selection. In fact, we will see later that we can almost always rule this model out over long time scales by looking at the actual parameter values of the model compared to what we know about species' population sizes and trait heritabilities.

Figure 3: Figure 6.3. Plot of species trait (x axis) versus fitness (y axis) showing a hypothetical landscape that would produce stabilizing selection. (figure stolen from web, need a better one!)

We can follow the modeling approach from chapter 3 to derive the expected distribution of species' traits on a tree under stabilizing selection. We can first consider the evolution of the trait on the "stem" branch, before the divergence of species A and B. We model stabilizing selection with a single optimal trait value, located at θ. An example of such a surface is plotted as Figure 6.3. We can describe fitness of an individual with phenotype z as:

(6.7)

$$W = e^{-\gamma(z-\theta)^2}$$

We have introduced a new variable, γ, which captures the curvature of the selection surface. To simplify the calculations, we will assume that stabilizing selection is weak, so that γ is small.

We can use a Taylor expansion of this function to approximate equation 6.7 using a polynomial. Our assumption that γ is small means that we can ignore terms of order higher than γ^2:

(6.8)

$$W = 1 - \gamma(z - \theta)^2$$

This makes good sense, since a quadratic equation is a good approximation of the shape of a normal distribution near its peak. The mean fitness in the population is then:

(6.9)

$$\begin{aligned}\bar{W} &= E[W] = E[1 - \gamma(z - \theta)^2] \\ &= E[1 - \gamma(z^2 - 2z\theta + \theta^2)] \\ &= 1 - \gamma(E[z^2] - E[2z\theta] + E[\theta^2]) \\ &= 1 - \gamma(\bar{z}^2 - V_z - 2\bar{z}\theta + \theta^2)\end{aligned}$$

We can find the rate of change of fitness by taking the derivative of (6.9) with respect to $\bar{z}$:

(6.10)

$$\frac{\partial \bar{W}}{\partial \bar{z}} = -2\gamma\bar{z} + 2\gamma\theta = 2\gamma(\theta - \bar{z})$$

We can now use Lande's (1976) equation for the dynamics of the population mean through time for a trait under selection:

(6.11)

$$\Delta\bar{z} = \frac{G}{\bar{W}}\frac{\partial \bar{W}}{\partial \bar{z}}$$

Substituting equations 6.9 and 6.10 into equation 6.11, we have:

(6.12)

$$\Delta\bar{z} = \frac{G}{\bar{W}}\frac{\partial \bar{W}}{\partial \bar{z}} = \frac{G}{1 - \gamma(\bar{z}^2 - V_z - 2\bar{z}\theta + \theta^2)} 2\gamma(\theta - \bar{z})$$

Then, simplifying further with another Taylor expansion, we obtain:

(6.13)

$$\bar{z}' = \bar{z} + 2G\gamma(\theta - \bar{z}) + \delta$$

Here, $\bar{z}$ is the species' trait value in the previous generation and $\bar{z}'$ in the next, while G is the additive genetic variance in the population, γ the curvature of the selection surface, θ the optimal trait value, and δ a random component capturing the effect of genetic drift. We can find the expected mean of the trait over time by taking the expectation of this equation:

(6.14)

$$E[\bar{z}'] = \mu_z' = \mu_z + 2G\gamma(\theta - \mu_z)$$

We can then solve this differential equations given the starting condition $\mu_z(0) = \bar{z}(0)$. Doing so, we obtain:

(6.15)

$$\mu_z(t) = \theta + e^{-2Gt\gamma}(\bar{z}(0) - \theta)$$

We can take a similar approach to calculate the expected variance of trait values. We use a standard expression for variance:

(6.16-18)

$$\begin{aligned} V_z' &= E[\bar{z}'^2] + E[\bar{z}']^2 \\ V_z' &= E[(\bar{z} + 2G\gamma(\theta - \bar{z}) + \delta)^2] - E[\bar{z} + 2G\gamma(\theta - \bar{z}) + \delta]^2 \\ V_z' &= G/n + (1 - 2G\gamma)^2 V_z \end{aligned}$$

If we assume that stabilizing selection is weak, we can simplify the above expression using a Taylor series expansion:

(6.19)

$$V_z' = G/n + (1 - 4G\gamma)V_z$$

We can then solve this differential equation with starting point $V_z(0) = 0$:

(6.20)

$$V_z(t) = \frac{e^{-4Gt\gamma} - 1}{4n\gamma}$$

This is equivalent to a standard stochastic model for constrained random walks called an Ornstein-Uhlenbeck process. Typical Ornstein-Uhlenbeck processes

have three parameters: the starting value ($\bar{z}(0)$), the optimum (θ), the drift parameter (σ^2), and a parameter describing the strength of constraints (α). In our parameterization, $\bar{z}(0)$ and θ are as given, $\alpha = 2G$, and $\sigma^2 = G/n$.

We now need to know how OU models behave when considered along the branches of a phylogenetic tree. In the simplest case, we can describe the joint distribution of two species, A and B, descended from a common ancestor, z. Expressions for trait values of species A and B are:

(6.21-22)

$$\bar{a}' = \bar{a} + 2G\gamma(\theta - \bar{a}) + \delta$$
$$\bar{b}' = \bar{b} + 2G\gamma(\theta - \bar{b}) + \delta$$

Expected values of these two equations give equations for the means:

(6.23-24)

$$\mu_a' = \mu_a + 2G\gamma(\theta - \mu_a)$$
$$\mu_b' = \mu_b + 2G\gamma(\theta - \mu_b)$$

We can solve this system of differential equations, given starting conditions $\mu_a(0) = \bar{a}_0$ and $\mu_b(0) = \bar{b}_0$:

(6.25-26)

$$\mu_a'(t) = \theta + e^{-2Gt\gamma}(bara_0 - \theta)$$
$$\mu_b'(t) = \theta + e^{-2Gt\gamma}(barb_0 - \theta)$$

However, we can also note that the starting value for both a and b is the same as the ending value for species z on the root branch of the tree. If we denote the length of that branch as t_1 then:

(6.27)

$$E[\bar{a}_0] = E[\bar{b}_0] = E[\bar{z}(t_1)] = e^{-2Gt_1\gamma}(\bar{z}_0 - \theta)$$

Substituting this into equations (6.25-26):

(6.28-29)

$$\mu_a'(t) = \theta + e^{-2G\gamma(t_1+t)}(\bar{z}_0 - \theta)$$
$$\mu_b'(t) = \theta + e^{-2G\gamma(t_1+t)}(\bar{z}_0 - \theta)$$

We can calculate the expected variance across replicates of species A and B, as above:

(6.30-32)

$$V'_a = E[\bar{a}'^2] + E[\bar{a}']^2$$
$$V'_a = E[(\bar{a} + 2G\gamma(\theta - \bar{a}) + \delta)^2] + E[\bar{a} + 2G\gamma(\theta - \bar{a}) + \delta]^2$$
$$V'_a = G/n + (1 - 2G\gamma)^2 V_a$$

Similarly,

(6.33-34)

$$V'_b = E[\bar{b}'^2] + E[\bar{b}']^2$$
$$V'_b = G/n + (1 - 2G\gamma)^2 V_b$$

Again we can assume that stabilizing selection is weak, and simplify these expressions using a Taylor series expansion:

(6.35-36)

$$V'_a = G/n + (1 - 4G\gamma) V_a$$
$$V'_b = G/n + (1 - 4G\gamma) V_b$$

We have a third term to consider, the covariance between species A and B due to their shared ancestry. We can use a standard expression for covariance to set up a third differential equation:

(6.37-39)

$$V'_{ab} = E[\bar{a}'\bar{b}'] + E[\bar{a}']E[\bar{b}']$$
$$V'_{ab} = E[(\bar{a} + 2G\gamma(\theta - \bar{a}) + \delta)(\bar{b} + 2G\gamma(\theta - \bar{b}) + \delta)] + E[\bar{a} + 2G\gamma(\theta - \bar{a}) + \delta]E[\bar{a} + 2G\gamma(\theta - \bar{a}) + \delta]$$
$$V'_{ab} = V_{ab}(1 - 2G\gamma)^2$$

We again use a Taylor series expansion to simplify:

(6.40)

$$V'_{ab} = -4V_{ab}G\gamma$$

Note that under this model the covariance between A and B decreases through time following their divergence from a common ancestor.

We now have a system of three differential equations. Setting initial conditions $V_a(0) = V_{a0}$, $V_b(0) = V_{b0}$, and $V_{ab}(0) = V_{ab0}$, we solve to obtain:

(6.41-43)

$$V_a(t) = \frac{1 - e^{-4G\gamma t}}{4n\gamma} + V_{a0}$$
$$V_b(t) = \frac{1 - e^{-4G\gamma t}}{4n\gamma} + V_{b0}$$
$$V_{ab}(t) = V_{ab0} e^{-4G\gamma t}$$

We can further specify the starting conditions by noting that both the variance of A and B and their covariance have an initial value given by the variance of z at time t_1:

(6.44)

$$V_{a0} = V_{ab0} = V_{ab0} = V_z(t_1) = \frac{e^{-4G\gamma t_1} - 1}{4n\gamma}$$

Substituting 6.44 into 6.41-43, we obtain:

(6.45-47)

$$\begin{aligned} V_a(t) &= \tfrac{e^{-4G\gamma(t_1+t)}-1}{4n\gamma} \\ V_b(t) &= \tfrac{e^{-4G\gamma(t_1+t)}-1}{4n\gamma} \\ V_{ab}(t) &= \tfrac{e^{-4G\gamma t}-e^{-4G\gamma(t_1+t)}}{4n\gamma} \end{aligned}$$

Under this model, the trait values follow a multivariate normal distribution; one can calculate that all of the other moments of this distribution are zero. Thus, the set of means, variances, and covariances completely describes the distribution of A and B. Also, as γ goes to zero, the selection surface becomes flatter and flatter. Thus at the limit as γ approaches 0, these equations are equal to those for Brownian motion (see chapter 4).

This quantitative genetic formulation – which follows Lande (1976) – is different from the typical parameterization of the OU model for comparative methods. We can obtain the "normal" OU equations by substituting $\alpha = 2G\gamma$ and $\sigma^2 = G/n$:

(6.48-50)

$$\begin{aligned} V_a(t) &= \tfrac{\sigma^2}{2\alpha}(e^{-2\alpha(t_1+t)} - 1) \\ V_b(t) &= \tfrac{\sigma^2}{2\alpha}(e^{-2\alpha(t_1+t)} - 1) \\ V_a b(t) &= \tfrac{\sigma^2}{2\alpha} e^{-2\alpha t}(1 - e^{-2\alpha t_1}) \end{aligned}$$

These equations are mathematically equivalent to the equations in Butler et al. (2004) applied to a phylogenetic tree with two species.

We can easily generalize this approach to a full phylogenetic tree with n taxa. In that case, the n species trait values will all be drawn from a multivariate normal distribution. The mean trait value for species i is then:

(6.51)

$$\mu_i(t) = \theta + e^{-2G\gamma T_i}(\bar{z}_0 - \theta)$$

Here T_i represents the total branch length separating that species from the root of the tree. The variance of species i is:

(6.52)

$$V_i(t) = \frac{e^{-4G\gamma T_i} - 1}{4n\gamma}$$

Finally, the covariance between species i and j is:

(6.53)

$$V_{ij}(t) = \frac{e^{-4G\gamma(T_i - s_{ij})} - e^{-4G\gamma T_i}}{4n\gamma}$$

Note that the above equation is only true when $T_i = T_j$ – which is only true for all i and j if the tree is ultrametric. We can substitute the normal OU parameters, α and σ^2, into these equations:

(6.54-56)

$$\mu_i(t) = \theta + e^{-\alpha T_i}(\bar{z}_0 - \theta)$$
$$V_i(t) = \frac{\sigma^2}{2\alpha} e^{-2\alpha T_i} - 1$$
$$V_{ij}(t) = \frac{\sigma^2}{2\alpha}(e^{-2\alpha(T_i - s_{ij})} - e^{-2\alpha T_i})$$

We can fit an OU model to data in a similar way to how we fit BM models in the previous chapters. For any given parameters ($\bar{z}_0$, σ^2, α, and θ) and a phylogenetic tree with branch lengths, one can calculate an expected vector of species means and a species variance-covariance matrix. One then uses the likelihood equation for a multivariate normal distribution to calculate the likelihood of this model. This likelihood can then be used for parameter estimation in either a ML or a Bayesian framework.

We can illustrate how this works by fitting an OU model to the mammal body size data that we have been discussing. Using ML, we obtain parameter estimates $\bar{z}_0 = 4.60$, $\sigma^2 = 0.10$, $\alpha = 0.0082$, and $\theta = 4.60$. This model has a lnL of -77.6, a little higher than BM, but an AIC_c score of 161.2, worse than BM. We still prefer Brownian motion for these data. Over many datasets, though, OU models fit better than Brownian motion (see Harmon et al. 2010, Pennell et al. 2015).

Section 6.6: Early burst models

Adaptive radiations are a slippery idea. Many definitions have been proposed, some of which contradict one another. Despite some core disagreement about the concept of adaptive radiations, many discussions of the phenomenon center

around the idea of "ecological opportunity." Perhaps adaptive radiations begin when lineages gain access to some previously unexploited area of niche space. These lineages begin diversifying rapidly, forming many and varied new species. At some point, though, one would expect that the ecological opportunity would be "used up," so that species would go back to diversifying at their normal, background rates. These ideas connect to Simpson's description of evolution in adaptive zones. According to Simpson, species enter new adaptive zones in one of three ways: dispersal to a new area, extinction of competitors, or the evolution of a new trait or set of traits that allow them to interact with the environment in a new way.

One idea, then, is that we could detect the presence of adaptive radiations by looking for bursts of trait evolution deep in the tree. If we can identify clades, like Darwin's finches, for example, that might be adaptive radiations, we should be able to uncover this "early burst" pattern of trait evolution.

The simplest way to model an early burst of evolution in a continuous trait is to use a time-varying Brownian motion model. Imagine that species in a clade evolved under a Brownian motion model, but one where the Brownian rate parameter (σ^2) slowed through time. In particular, we can follow Harmon et al. (2010) and define the rate parameter as a function of time, as:

(6.57)

$$\sigma^2(t) = \sigma_0^2 e^{rt}$$

We describe the rate of decay of the rate using the parameter r, which must be negative to fit our idea of adaptive radiations. The rate of evolution will slow through time, and will decay more quickly if the absolute value of r is large.

This model also generates a multivariate normal distribution of tip values. Harmon et al. (2010) followed Blomberg's "ACDC" model to write equations for the means and variances of tips on a tree under this model, which are:

(6.58-60)

$$\mu_i(t) = \bar{z}_0$$
$$V_i(t) = \sigma_0^2 \frac{e^{rT_i}-1}{r} V_{ij}(t) = \sigma_0^2 \frac{e^{rs_{ij}}-1}{r}$$

Again, we can generate a vector of means and a variance-covariance matrix for this model given parameter values ($\bar{z}_0$, σ^2, and r) and a phylogenetic tree. We can then use the multivariate normal probability distribution function to calculate a likelihood, which we can then use in a ML or Bayesian statistical framework.

For mammal body size, the early burst model does not explain patterns of body size evolution, at least for the data considered here ($\bar{z}_0 = 4.64$, $\sigma^2 = 0.088$, $r = -0.000001$, $lnL = -78.0$, $AIC_c = 162.6$).

Section 6.7: Peak shift models

A second model considered by Hansen and Martins (1996) describes the circumstance where traits change in a punctuated manner. One can imagine a scenario where species evolve on an adaptive landscape with many peaks; usually, populations stay on a single peak and phenotypes do not change, but occasionally a population will transition from one peak to another. We can either assume that these changes occur at random times, with an average interval between peak shifts of , or we can associate shifts with other traits that we map on the phylogenetic tree (for example, major geographic dispersal or vicariance events, or the evolution of certain traits.

We have developed peak shift models by integrating OU models and reversible-jump MCMC (Uyeda et al. 2014). The mathematics of this model are beyond the scope of this book, but follow closely from the description of the multi-rate Brownian motion model described in the section "variation in rates of trait evolution across clades," above. In this case, when we change model parameters, we move among OU regimes, and can alter any of the OU model parameters (or). The approach can be used to either identify parts of the tree that are evolving in separate regimes or to test particular hypotheses about the drivers of evolution.

Section 6.8: Summary

In this chapter, I have described a few models that represent alternatives to Brownian motion, which is still the dominant model of trait evolution used in the literature. These examples really represent the beginnings of a whole set of models that one might fit to biological data. The best applications of this type of approach, I think, are in testing particular biologically motivated hypotheses using comparative data.

References

Alfaro, M. E. 2013. VI.15. key evolutionary innovations. *in* J. B. Losos, D. A. Baum, D. J. Futuyma, H. E. Hoekstra, R. E. Lenski, A. J. Moore, C. L. Peichel, D. Schluter, and M. C. Whitlock, eds. The princeton guide to evolution. Princeton University Press, Princeton.

Blomberg, S. P., T. Garland Jr, and A. R. Ives. 2003. Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. Evolution 57:717–745.

Boettiger, C., G. Coop, and P. Ralph. 2012. Is your phylogeny informative? Measuring the power of comparative methods. Evolution 66:2240–2251.

Bokma, F. 2008. DETECTION OF "PUNCTUATED EQUILIBRIUM" BY

BAYESIAN ESTIMATION OF SPECIATION AND EXTINCTION RATES, ANCESTRAL CHARACTER STATES, AND RATES OF ANAGENETIC AND CLADOGENETIC EVOLUTION ON a MOLECULAR PHYLOGENY. Evolution 62:2718–2726. Blackwell Publishing Inc.

Eastman, J. M., M. E. Alfaro, P. Joyce, A. L. Hipp, and L. J. Harmon. 2011. A novel comparative method for identifying shifts in the rate of character evolution on trees. Evolution 65:3578–3589.

Garland, T., Jr. 1992. Rate tests for phenotypic evolution using phylogenetically independent contrasts. Am. Nat. 140:509–519.

Goldberg, E. E., and B. Igić. 2012. Tempo and mode in plant breeding system evolution. Evolution 66:3701–3709. Wiley Online Library.

Grant, P. R., and B. R. Grant. 2002. Unpredictable evolution in a 30-year study of darwin's finches. Science 296:707–711.

Grant, P. R., and B. Rosemary Grant. 2011. How and why species multiply: The radiation of darwin's finches. Princeton University Press.

Harmon, L. J., J. B. Losos, T. Jonathan Davies, R. G. Gillespie, J. L. Gittleman, W. Bryan Jennings, K. H. Kozak, M. A. McPeek, F. Moreno-Roark, T. J. Near, and Others. 2010. Early bursts of body size and shape evolution are rare in comparative data. Evolution 64:2385–2396.

Hunter, J. P. 1998. Key innovations and the ecology of macroevolution. Trends Ecol. Evol. 13:31–36.

Lande, R. 1976. Natural selection and random genetic drift in phenotypic evolution. Evolution 30:314–334.

O'Meara, B. C., C. Ané, M. J. Sanderson, and P. C. Wainwright. 2006. Testing for different rates of continuous trait evolution using likelihood. Evolution 60:922–933.

Pagel, M. 1999a. Inferring the historical patterns of biological evolution. Nature 401:877–884.

Pagel, M. 1999b. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. Syst. Biol. 48:612–622. [Oxford University Press, Society of Systematic Biologists].

Revell, L. J. 2013. Two new graphical methods for mapping trait evolution on phylogenies. Methods Ecol. Evol. 4:754–759.

Revell, L. J., L. J. Harmon, and D. C. Collar. 2008. Phylogenetic signal, evolutionary process, and rate. Syst. Biol. 57:591–601. Oxford University Press.

Thomas, G. H., R. P. Freckleton, and T. Székely. 2006. Comparative analyses of the influence of developmental mode on phenotypic diversification rates in

shorebirds. Proc. Biol. Sci. 273:1619–1624.

Uyeda, J. C., and L. J. Harmon. 2014. A novel bayesian method for inferring and interpreting the dynamics of adaptive landscapes from phylogenetic comparative data. Syst. Biol. 63:902–918.

Yoder, J. B., E. Clancey, S. Des Roches, J. M. Eastman, L. Gentry, W. Godsoe, T. J. Hagey, D. Jochimsen, B. P. Oswald, J. Robertson, and Others. 2010. Ecological opportunity and the origin of adaptive radiations. J. Evol. Biol. 23:1581–1596. Blackwell Publishing Ltd.

Chapter 7: Models of discrete character evolution

Biological motivation: Limblessness as a discrete trait

Squamates, the clade that includes all living species of lizards, are well known for their diversity. From the gigantic Komodo dragon of Indonesia (Figure 7.1A, *Varanus komodoensis*) to tiny leaf chameleons of Madagascar (Figure 7.1B, *Brookesia*), squamates span an impressive range of form and ecological niche use. Even the snakes (Figure 7.1C and D), extraordinarily diverse in their own right (~3,500 species), are actually a clade that is nested within squamates. The squamate lineage that is ancestral to snakes became limbless about 170 million years ago (see Timetree of Life) – and also underwent a suite of changes to their head shape, digestive tract, and other traits associated with their limbless lifestyle. In other words, snakes are lizards – highly modified lizards, but lizards nonetheless. And snakes are not the only limbless lineage of squamates. In fact, lineages within squamates have lost their limbs over and over again through their history (e.g. Figure 7.1E and F), with some estimates that squamates have lost their limbs at least 26 times in the past 240 million years.

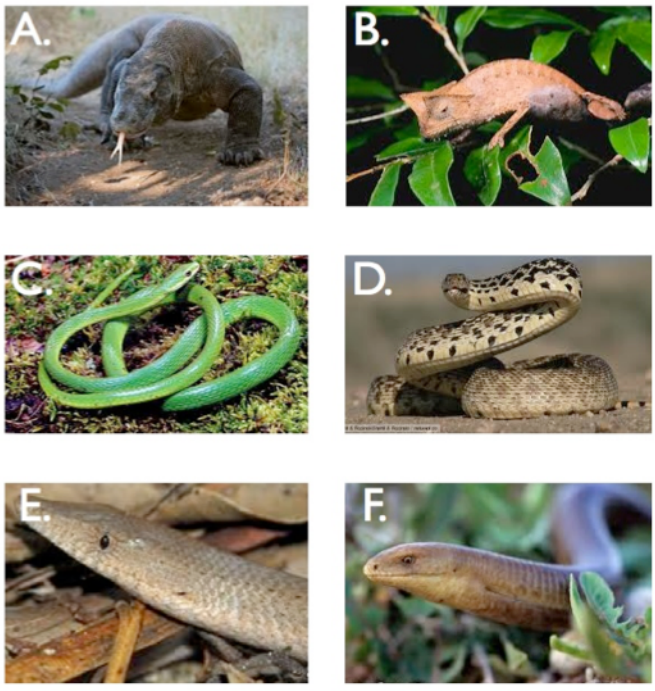

Figure 1: Figure 7.1. Squamates, legged and legless. A. Komodo dragon, B. Brookesia chameleon, C. and D. snakes, E. and F. legless lizards.

Limblessness is an example of a discrete trait – a trait that can occupy one of a set of distinct character states. Analyzing the evolution of discrete traits requires a different modeling approach than what we used for continuous traits. In this chapter, I will introduce the Mk model, which is a general approach to modeling the evolution of discrete traits on trees. Fitting this model to comparative data will help us understand the evolution of traits like limblessness where species can be placed into one of a number of discrete character states.

Key Questions

- How can we model the evolution of discrete characters that have a set number of fixed states?
- How can we change the parameters of the Mk model to construct more elaborate models of discrete character evolution?
- How do we simulate the evolution of a discrete character?

Modeling the evolution of discrete states

So far, we have only dealt with continuously varying characters. However, many characters of interest to biologists can best be described by characters with a set number of fixed states. For limblessness in squamates, introduced above (snakes and lizards), each species is either legless (state 0) or not (state 1; actually, there are some species that might be considered "intermediate," but we will ignore that here). We might have particular questions about the evolution of limblessness in squamates. For example, how many times this character has changed in the evolutionary history of squamates? How often does limblessness evolve? Do limbs ever re-evolve? Is the evolution of limblessness related to some other aspect of the lives of these reptiles?

We will consider discrete characters where each species might exhibit one of k states. (In the limbless example above, k=2). For characters with more than two states, there is a key distinction between ordered and unordered characters. Ordered characters can be placed in an order so that transitions only occur between adjacent states. For example, I might include "intermediate" species that are somewhere in between limbed and limbless – for example, the "mermaid skinks" (Sirenoscincus) from Madagascar, so called because they lack hind limbs (Figure 7.2). An ordered model might only allow transitions between limbless and intermediate, and intermediate and limbed; it would be impossible under such a model to go directly from limbed to limbless without first becoming intermediate. For unordered characters, any state can change into any other state. In this chapter, I will focus mainly on unordered characters; we will return to ordered characters later in the book.

Most work on the evolution of discrete characters on phylogenetic trees has focused on the evolution of gene or protein sequences. Gene sequences are made up of four character states (A, C, T, and G for DNA). Models of sequence

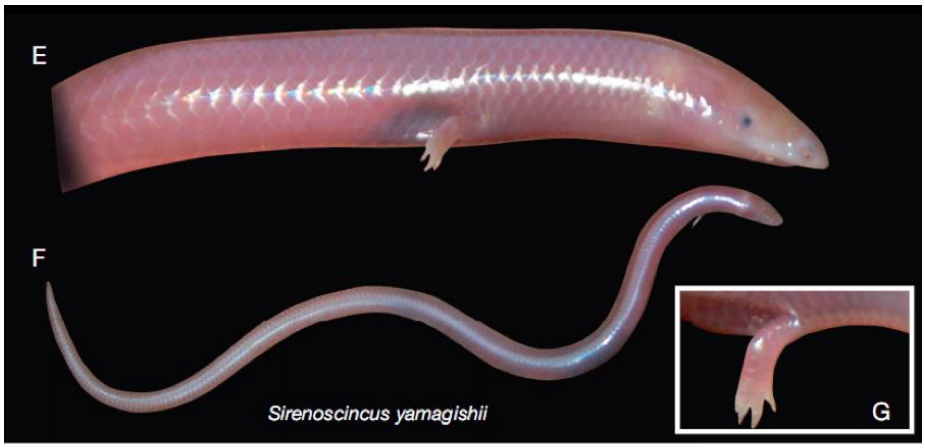

FIG. 1. — **E-G**, living specimen of *Sirenoscincus yamagishii* Sakata & Hikida, 2003 from Ankarafantsika, Madagascar, lateral view of the anterior body part (**E**), dorsolateral view of the entire specimen (**F**), and close-up of the right forelimb with four claws (**G**). (Photographs E-G: Falk S. Eckhardt.)

Figure 2: Figure 7.2. Mermaid skink

evolution allow transitions among all of these states at certain rates, and may allow transition rates to vary across sites, among clades, or through time. There are a huge number of named models that have been applied to this problem (e.g. Jukes-Cantor, JC; General Time-Reversible, GTR; and many more), and a battery of statistical approaches are available to fit these models to data (e.g. Posada and Crandall 1998).

Any discrete character can be modeled in a similar way as gene sequences. When considering phenotypic characters, we should keep in mind two main differences from the analysis of DNA sequences. First, arbitrary discrete characters may have any number of states (beyond the four associated with DNA sequence data). Second, characters are typically analyzed independently rather than combining long sets of characters and assuming that they share the same model of change.

The Mk Model

The most basic model for discrete character evolution is called the Mk model. First developed for trait data by Pagel (1994; although the name Mk comes from Lewis 2001), this model is a direct analogue of the Jukes-Cantor (JC) model for sequence evolution. The model applies to a discrete character having k unordered states. Such a character might have $k = 2$, $k = 3$, or even more states. Evolution involves changing between these k states (Figure 7.3).

The basic version of the Mk model assumes that transitions among these states follow a Markov process. This means that the probability of changing from one state to another depends only on the current state, and not on what has come

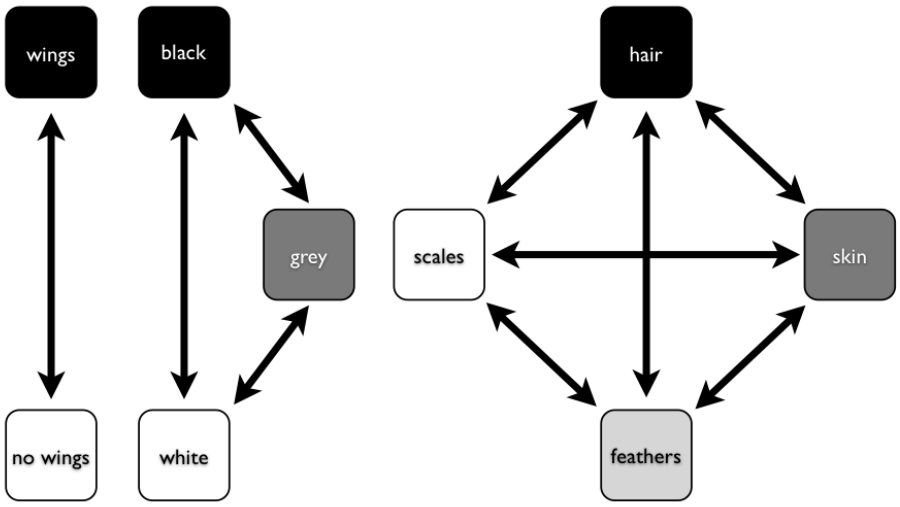

Figure 3: Figure 7.3. Examples of discrete characters with (A) $k = 2$, (B) $k = 3$, and (C) $k = 4$ states.

before. For example, it makes no difference if a lineage has just evolved the trait of "feathers," or whether they have had feathers for millions of years – the probability of evolving a different character state is the same in both cases. The basic Mk model also assumes that every state is equally likely to change to any other states.

For the basic Mk model, we can denote the instantaneous rate of change between states using the parameter q. In general, q_{ij} is called the instantaneous rate between character states i and j. It is defined as the limit of the rate measured over very short time intervals. Imagine that you calculate a rate of character change by counting the number of changes of state of a character over some time interval, t. The instantaneous rate is the value that this rate approaches as t gets smaller and smaller so that the time interval is nearly zero. Again, for the basic Mk model, instantaneous rates between all pairs of characters are equal; that is, $q_{ij} = q_{mn}$ for all $i \neq j$ and $m \neq n$.

We can summarize general Markov models for discrete characters using a transition rate matrix:

(eq. 7.1)

$$\mathbf{Q} = \begin{bmatrix} -r_1 & q_{12} & \dots & q_{1k} \\ q_{21} & -r_2 & \dots & q_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ q_{k1} & qk2 & \dots & -r_k \end{bmatrix}$$

Note that the instantaneous rates are only entered into the off-diagonal parts of the matrix. Along the diagonal, these matrices always have a set of negative numbers. For any $\mathbf{Q}$ matrix, the sum of all the elements in each row is zero – a necessary condition for a transition rate matrix. Because of this, each negative number has a value, r_i, equal to the sum of all of the other numbers in the row. For example,

(eq. 7.2)

$$r_1 = \sum_{i=2}^{k} q_{1i}$$

For a two-state Mk model, $k = 2$ and rates are symmetric so that $q_{12} = q_{21}$. In this case, we can write the transition rate matrix as:

(eq. 7.3)

$$\mathbf{Q} = \begin{bmatrix} -q & q \\ q & -q \end{bmatrix}$$

Likewise, for $k = 3$,

(eq. 7.4)

$$\mathbf{Q} = \begin{bmatrix} -2q & q & q \\ q & -2q & q \\ q & q & -2q \end{bmatrix}$$

In general,

(eq. 7.5)

$$\mathbf{Q} = q \begin{bmatrix} 1-k & 1 & \dots & 1 \\ 1 & 1-k & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix}$$

Once we have this transition rate matrix, we can calculate the probability distribution of trait states after any time interval t using the equation:

(eq. 7.6)

$$\mathbf{P}(t) = e^{\mathbf{Q}t}$$

This equation looks simple, but calculating $P(t)$ involves matrix exponentiation – raising e to a power defined by a matrix. This calculation is substantially different from raising e to the power defined by each element of a matrix. I will not cover the details of matrix exponentiation here – interested readers should see Yang (2006) for details – but the calculations are not trivial. The result is a matrix, $\mathbf{P}$, of transition probabilities. Each element in this matrix (p_{ij}) gives the probability that starting in state i you will end up in state j over that time interval t. For the standard Mk model, there is a general solution to this equation:

(eq. 7.7)

$$\begin{aligned} p_{ii}(t) &= \tfrac{1}{k} + \tfrac{k-1}{k} e^{-kqt} \\ p_{ij}(t) &= \tfrac{1}{k} - \tfrac{1}{k} e^{-kqt} \end{aligned}$$

In particular, when k = 2,

(eq. 7.8)

$$\begin{aligned} p_{ii}(t) &= \tfrac{1}{k} + \tfrac{k-1}{k} e^{-kqt} = \tfrac{1}{2} + \tfrac{2-1}{2} e^{-2qt} = \tfrac{1+e^{-2qt}}{2} \\ p_{ij}(t) &= \tfrac{1}{k} - \tfrac{1}{k} e^{-kqt} = \tfrac{1}{2} - \tfrac{1}{2} e^{-2qt} = \tfrac{1-e^{-2qt}}{2} \end{aligned}$$

If we consider what happens when time gets very large in these equations, we see an interesting pattern. Any term that has e^{-t} in it gets closer and closer to

zero as t increases. Because of this, for all values of k, each $p_{ij}(t)$ converges to a constant value, $1/k$. This is the stationary distribution of character states, π, defined as the equilibrium frequency of character states if the process is run many times for a long enough time period. In general, the stationary distribution of an Mk model is:

(eq. 7.9)

$$\pi = \begin{bmatrix} 1/k & 1/k & \dots & 1/k \end{bmatrix}$$

In the case of k = 2,

(eq. 7.10)

$$\pi = \begin{bmatrix} 1/2 & 1/2 \end{bmatrix}$$

The Extended Mk Model

The Mk model assumes that transitions among all possible character states occur at the same rate. However, that may not be a valid assumption. For example, it is often supposed that it is easier to lose a complex character than to gain one. We might want to fit models that allow for such asymmetries in rates.

For models of DNA sequence evolution there are a wide range of models allowing different rates between distinct types of nucleotides. Unequal rates are usually incorporated into the Mk model in two ways. First, one can consider the symmetric model (SYM). In the symmetric model, the rate of change between any two character states is the same forwards as it is backwards (that is, rates of change are symmetric; $q_{ij} = q_{ji}$). The rate for a particular pair of states might differ from other pairs of character stats. The rate matrix for this model has as many free rate parameters as there are pairs of character states, $k(k-1)/2$.

(eq. 7.11)

$$p = \frac{k(k-1)}{2}$$

However, in general symmetric models will not have stationary distributions where all character states occur at equal frequencies, as noted above for the Mk model. We can account for these uneven frequencies by adding additional parameters to our model:

(eq. 7.12)

$$\pi_{SYM} = \begin{bmatrix} \pi_1 & \pi_2 & \dots & 1 - \sum_{i=1}^{n-1} \pi_i \end{bmatrix}$$

Note that we only have to specify $n-1$ equilibrium frequencies, since we know that they all sum to one. We have added $n-1$ new parameters, for a total number of parameters:

(eq. 7.13)

$$p = \frac{k(k-1)}{2} + n - 1$$

To obtain a **Q**-matrix for this model, we combine the information from both the relative transition rates and equilibrium frequencies:

(eq. 7.14)

$$\mathbf{Q} = \begin{bmatrix} \cdot & r_1 & \dots & r_{n-1} \\ r_1 & \cdot & \dots & \vdots \\ \vdots & \vdots & \cdot & r_{k(k-1)/2} \\ r_{n-1} & \dots r_{k(k-1)/2} & \cdot & \end{bmatrix} \begin{bmatrix} \pi_1 & 0 & 0 & 0 \\ 0 & \pi_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \pi_n \end{bmatrix}$$

In this equation I have left the diagonal of the first matrix as dots. The final **Q**-matrix must have all rows sum to one, so one can adjust the values of that matrix after the multiplication step.

In the case of a two-state model, for example, we can create a model where the forward rate is double the backward rate, and the equilibrium frequency of character one is 0.75. Then:

(eq. 7.15)

$$\mathbf{Q} = \begin{bmatrix} \cdot & 1 \\ 2 & \cdot \end{bmatrix} \begin{bmatrix} 0.75 & 0 \\ 0 & 0.25 \end{bmatrix} = \begin{bmatrix} \cdot & 0.25 \\ 1.5 & \cdot \end{bmatrix} = \begin{bmatrix} -0.25 & 0.25 \\ 1.5 & -1.5 \end{bmatrix}$$

The second common extension of the Mk model is called the all-rates-different model (ARD). In this model every possible type of transition can have a different rate. There are thus $k(k-1)$ free rate parameters for this model, and again $n-1$ parameters to specify the equilibrium frequencies of the character states.

The same algorithm can be used to calculate the likelihood for both of these extended Mk models (SYM and ARD). These models have more parameters than the standard Mk. To find maximum likelihood solutions, we must optimize the likelihood across the entire set of unknown parameters (see Chapter 7).

Simulating the Mk model on a tree

We can also use the equations above to simulate evolution under an Mk or extended-Mk model on a tree. To do this, we simulate character evolution on each branch of the tree, starting at the root and progressing towards the tips. At speciation, we assume that both daughter species inherit the character state of their parental species immediately following speciation, and then evolve independently after that. At the end of the simulation, we will obtain a set of character states, one for each tip in the tree. The distribution of character states will depend on the shape of the phylogenetic tree (both its topology and branch lengths) along with the parameters of our model of character evolution.

We first draw a beginning character state at the root of the tree. There are several common ways to do this. For example, we can either draw from the stationary distribution or from one where each character state is equally likely. In the case of the standard Mk model, these are the same. For example, if we are simulating evolution under Mk with $k = 2$, then state 0 and 1 each have a probability of 1/2 at the root. We can draw the root state from a binomial distribution with $p = 0.5$.

Once we have a character state for the root, we then simulate evolution along each branch in the tree. We start with the (usually two) branches descending from the root. We then proceed up the tree, branch by branch, until we get to the tips.

We can understand this algorithm perfectly well by thinking about what happens on each branch of the tree, and then extending that algorithm to all of the branches (as described above). For each branch, we first calculate $\mathbf{P}(t)$, the transition probability matrix, given the length of the branch and our model of evolution as summarized by $\mathbf{Q}$ and the branch length t. We then focus on the row of $\mathbf{P}(t)$ that corresponds to the character state at the beginning of the branch. For example, let's consider a basic two-state Mk model with $q = 0.5$. We will call the states 0 and 1. We can calculate $\mathbf{P}(t)$ for a branch with length $t = 3$ as:

(eq. 7.16)

$$\mathbf{P}(t) = e^{\mathbf{Q}t} = exp(\begin{bmatrix} -0.5 & 0.5 \\ 0.5 & -0.5 \end{bmatrix} \cdot 3) = \begin{bmatrix} 0.525 & 0.475 \\ 0.475 & 0.525 \end{bmatrix}$$

If we had started with character state 0 at the beginning of this branch, we would focus on the first row of this matrix. We want to end up at state 0 with probability 0.525 and change to state 1 with probability 0.475. We again draw a uniform random deviate u, and choose state 0 if $0 \leq u < 0.525$ and state 1 if $0.525 \leq u < 1$. If we started with a different character state, we would use a different row in the matrix. If this is an internal branch in the tree, then both daughter species inherit the character state that we chose immediately following

speciation – but might diverge soon after! By repeating this along every branch in the tree, we obtain a set of character states at the tips of the tree. This is then the output of our simulation.

Two additional details here are worth noting. First, the procedure for simulating characters under the extended-Mk model are identical to those above, except matrix exponentials are more complicated than in the standard Mk model. Second, if you are simulating a character with more than two states, then the procedure for drawing a random number is slightly different. One still obtains the relevant row from $\mathbf{P}(t)$ and draws a uniform random deviate u. Imagine that we have a ten-state character with states 0 - 9. We start at state 0 at the beginning of the simulation. Again using $q = 0.5$ and $t = 3$, we find that:

(eq. 7.17)

$$\begin{aligned} p_{ii}(t) &= \tfrac{1}{k} + \tfrac{k-1}{k} e^{-kqt} = \tfrac{1}{10} + \tfrac{9}{10} e^{-2 \cdot 0.5 \cdot 3} = 0.145 \\ p_{ij}(t) &= \tfrac{1}{k} - \tfrac{1}{k} e^{-kqt} \tfrac{1}{10} - \tfrac{1}{10} e^{-2 \cdot 0.5 \cdot 3} = 0.095 \end{aligned}$$

We focus on the first row of $\mathbf{P}(t)$, which has elements:

$$\begin{bmatrix} 0.145 & 0.095 & 0.095 & 0.095 & 0.095 & 0.095 & 0.095 & 0.095 & 0.095 & 0.095 \end{bmatrix}$$

We calculate the cumulative sum of these elements, adding them together so that each number represents the sum of itself and all preceding elements in the vector:

$$\begin{bmatrix} 0.145 & 0.240 & 0.335 & 0.430 & 0.525 & 0.620 & 0.715 & 0.810 & 0.905 & 1.000 \end{bmatrix}$$

Now we compare u to the numbers in this cumulative sum vector. We select the smallest element that is still strictly larger than u, and assign this character state for the end of the branch. For example, if $u = 0.475$, the 5th element, 0.525, is the smallest number that is still greater than u. This corresponds to character state 4, which we assign to the end of the branch. This last procedure is a numerical trick. Imagine that we have a line segment with length 1. The cumulative sum vector breaks the unit line into segments, each of which is exactly as long as the probability of each event in the set. One then just draws a random number between 0 and 1 using a uniform distribution. The segment that contains this random number is our event.

We can apply this approach to simulate the evolution of limblessness in squamates. Below, I present the results of three such simulations. These simulations are a little different than what I describe above because they consider all changes in the tree, rather than just character states at nodes and tips; but the model (and the principal) is the same. You can see that the model leaves an imprint

on the pattern of changes in the tree, and you can imagine that one might be able to reconstruct the model using a phylogenetic comparative approach. Of course, typically we know only the tip states, and have to reconstruct changes along branches in the tree. We will discuss parameter estimation for the Mk and extended-Mk models in the next chapter.

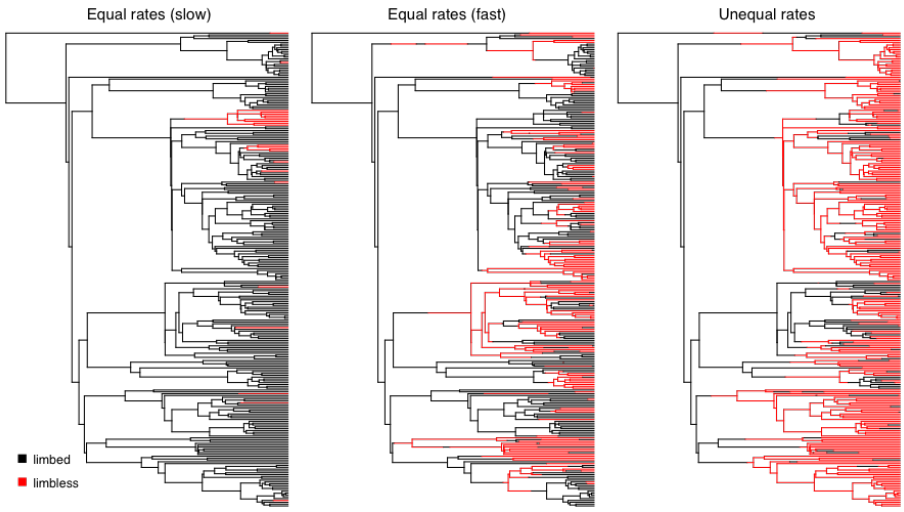

Figure 4: Figure 7.4. Simulated character evolution on a phylogenetic tree of squamates (from Brandley et al. 2008) under an equal-rates Mk model with slow, fast, and asymmetric transition rates (from right to left). In all three cases, I assumed that the ancestor of squamates had limbs.

Chapter summary

In this chapter I have described the Mk model, which can be used to describe the evolution of discrete characters that have a set number of fixed states. We can also elaborate on the Mk model to allow more complex models of discrete character evolution (the extended-Mk model). These models can all be used to simulate the evolution of discrete characters on trees.

In summary, the Mk and extended Mk model are general models that one can use for the evolution of discrete characters. In the next chapter, I will show how to fit these models to data and use them to test evolutionary hypotheses.

Chapter 8: Fitting models of discrete character evolution

Biological motivation: The evolution of limbs and limblessness

In the introduction to Chapter 7, I mentioned that squamates had lost their limbs repeatedly over their evolutionary history. This is a pattern that has been known for decades, but analyses have been limited by the lack of a well-supported species-level phylogenetic tree of squamates. Only in the past few years have phylogenetic trees been produced at a scale broad enough to take a comprehensive look at this question (e.g. Bergmann et al. 2011; Pyron et al. 2013; see Figure 8.1). Such efforts to reconstruct this section of the tree of life provide exciting potential to revisit old questions with new data.

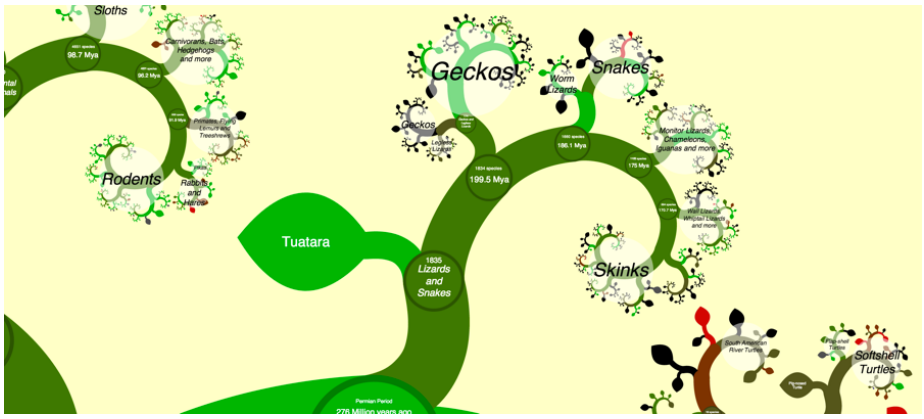

Figure 1: Figure 8.1. A view of the squamate tree of life. Data from Bergmann et al. 2011, visualized using OneZoom (Rosindell and Harmon 2012; see www.onezoom.org).

Plotting the pattern of limbed and limbless species on the tree leads to interesting questions about the tempo and mode of this trait in squamates. For example, are there multiple gains as well as losses of limbs? Do gains and losses happen at the same rate, or (as we might expect) are gains more rare than losses? We can test hypothesis such as these using the the Mk and extended-Mk models (see chapter 7). In this chapter we will fit these models to phylogenetic comparative data.

Key Biological Questions

- How do we calculate the likelihoods of Mk and extended-Mk models on phylogenetic trees?
- How can we use these approaches to test hypotheses about character evolution?

Fitting Mk models to comparative data

The equations in Chapter 7 give us enough information to calculate the likelihood for comparative data on a tree. To understand how this is done, we can start with the simplest case, where we know the beginning state of a character, the branch length, and the end state. We can then apply this method across an entire tree using a pruning algorithm, which will allow calculation of the likelihood of the data given the model and phylogenetic tree.

Imagine that a two-state character changes from a state of 0 to a state of 1 sometime over a time interval of t = 3. We can set q = 0.5 to calculate a probability matrix:

(eq. 8.1)

For this simple example, we started with state 0, so we look at the first row. We ended at state 1, so we look specifically at P12(t): the probability of starting with state 0 and ending with state 1 over time t. This is the probability of obtaining the data given the model (i.e. the likelihood): L = 0.475.

When we have comparative data the situation is more complex. If we knew the ancestral character states and states at every node in the tree, then calculation of the overall likelihood would be straightforward – we could just apply the approach above many times, once for each branch of the tree. However, there are two problems with this approach. First, we don't know the starting state of the character at the root of the tree, and must treat that as an unknown. Second, we are modeling a process that is happening independently on many branches in a phylogenetic tree, and only observe the states at the end of these branches. All of the character states at internal nodes of the tree are unknown. The likelihood that we want to calculate has to be summed across all of these unknown character state possibilities on the internal branches of the tree.

Thankfully, Felsenstein (1973) provides an elegant algorithm for calculating the likelihoods for discrete characters on a tree. This algorithm, called Felsenstein's pruning algorithm, is described with an example in box 8.1. Felsenstein's pruning algorithm was important in the history of phylogenetics because it allowed scientists to efficiently calculate the likelihoods of comparative data given a tree and a model. One can then maximize that likelihood by changing model parameters (and perhaps also the topology and branch lengths of the tree; see Felsenstein 2003).

Box 8.1: Felsenstein's pruning algorithm

Felsenstein's pruning algorithm is an example of dynamic programming, a type of algorithm that has many applications in comparative biology. In dynamic programming, we break down a complex problem into a series of simpler steps that have a nested structure. This allows us to reuse computations in an efficient way and speeds up the time required to make calculations.

<<< Figure 8.2 (four panels) >>>

Panel A:

Panel B:

Panel C:

Panel D:

The best way to illustrate Felsenstein's algorithm is through an example, which is presented in the figure above. We are trying to calculate the likelihood for a three-state character on a phylogenetic tree that includes six species. In the figure, each tip and internal node in the tree has four boxes, which will contain the probabilities for the three character states at that point in the tree (Figure 8.2 A).

1. The first step in the algorithm is to fill in the probabilities for the tips. In this case, we know the states at the tips of the tree, so we just put a one in the box that corresponds to the actual character state and zeros in all others (Figure 8.2 B). In other words, we are stating that we know precisely the character states at the tips; the probability that that species has the state that we observe is 1, and all other states have probability zero.

2. Next, we identify a node where all of its immediate descendants are tips. There will always be at least one such node; often, there will be more than one, in which case we will arbitrarily choose one. For this example, we will choose the node that is the most recent common ancestor of species A and B, labeled as node 1 in Figure 8.2 B.

3. We then use equation 7.6 to calculate the conditional likelihood for each character state for the subtree that includes the node of interest and its tip descendants. For each character state, the conditional likelihood is the probability, given the data and the model, of obtaining the tip character states if you start with that character state at the root. In other words, we keep track of the likelihood for the tipward parts of the tree, including our data, if the node we are considering had each of the possible character states. This calculation is:

For the example given, the two tip character states are 0 (for species A) and 1 (for species B). We can calculate the conditional likelihood for character state 0 at node 1 as:

Notice that for x = 0 and for x = 1 and x = 2; similarly, for x = 1 and for x = 0 and x = 2. We can calculate the probability terms from the probability matrix. In this case , so for both the left and right branch:

so that

and

finally

We can use a similar approach to find that:

and

These numbers are entered into the appropriate boxes in Figure xxx.C.

4. We then repeat the above calculation for every node in the tree. For nodes 3-5, not all of the and terms are zero; their values can be read out of the boxes on the tree. The result of all of these calculations can be seen in Figure 8.2 D.
5. We can now calculate the likelihood across the whole tree using the conditional likelihoods for the three states at the root of the tree. We use the equation:

Where is the prior probability of that character state at the root of the tree. We will take these prior probabilities to be equal for each state, or uniform (); two other possibilities are given in the text. The likelihood for our example, then, is:

Note that if you try this example in another software package, like GEIGER or PAUP*, the software will calculate a ln-likelihood of -6.5, which is exactly the natural log of the value calculated here.

Felsenstein's pruning algorithm proceeds backwards in time from the tips to the root of the tree (see Box 8.1). At the root, we must specify the probabilities of each character state in the common ancestor of the species in the clade. As mentioned in Chapter 7, there are at least three possible methods for doing this. First, one can assume that each state can occur at the root with equal probability. Second, one can assume that the states are drawn from their stationary distribution, as given by the model. The stationary distribution is a stable probability distribution of states that is reached by the model after a long amount of time. Third, one might have some information about the root state – perhaps from fossils, or information about character states in a set of outgroup taxa – that can be used to assign probabilities to the states. In practice, the first two of these methods are more common. In the case discussed above – an Mk model with all transition rates equal – the stationary distribution is one where

all states are equally probable, so the first two methods are identical. In general, though, these three methods can give different results.

Using maximum likelihood to estimate parameters of the Mk model

The algorithm in Box 8.1 gives the likelihood for any particular discrete-state Markov model on a tree, but requires us to specify a value of the rate parameter q. In the example given, this rate parameter q = 1.0 corresponds to a lnL of -6.5. But is this the best value of q to use for our Mk model? Probably not. We can use maximum likelihood to find a better estimate of this parameter.

If we apply the pruning algorithm across a range of different values of q, the likelihood changes. To find the ML estimate of q, we simply need to try a range of q-values, and stop at the value of q that has the highest log-likelihood.

The process of trying a range of possibilities for q is inefficient, though. A better strategy involves the use of optimization algorithms, a well-developed field of mathematical analysis. These algorithms differ in their details, but we can illustrate how they work with a general example. Imagine that you are near Mt. St. Helens, and you are tasked with finding the peak of that mountain. It is foggy, but you can see the area around your feet and have an accurate altimeter. One strategy is to simply look at the slope of the mountain where you are standing, and climb uphill. If the slope is steep, you probably still are far from the top, and should climb fast; if the slope is shallow, you might be near the top of the mountain. It may seem obvious that this will get you to a local peak, but perhaps not the highest peak of Mt. St. Helens. Mathematical optimization schemes have this potential difficulty as well, but use some tricks to jump around in parameter space and try to find the highest peak as they climb. Details of actual optimization algorithms are beyond the scope of this book; for more information, see Nocedal and Wright (2000).

XXX Lizard example

The example above considers maximization of a single parameter, which is a relatively simple problem. When we extend this to a multi-parameter model – for example, the extended Mk model will all rates different (ARD) – maximizing the likelihood becomes much more difficult. A large number of algorithms exist to solve this problem (multivariate optimization methods); this is a bit outside the scope of this chapter. The R exercise associated with this chapter will allow you to explore some R functions for optimization.

We can also analyze this model using a Bayesian MCMC framework. We can modify the standard approach to Bayesian MCMC (see chapter 2):

1. Sample a starting parameter value, q, from its prior distributions. For this example, we can set our prior distribution as uniform between 0 and

1. (Note that one can also treat probabilities of states at the root as a parameter to be estimated from the data).

2. Given the current parameter value, select new proposed parameter values using the proposal density . For example, we might use a uniform proposal density with width 0.2, so that .

3. Calculate three ratios:

a. The prior odds ratio. In this case, since our prior is uniform, this is 1.
b. The proposal density ratio. In this case our proposal density is symmetrical, so.

c. The likelihood ratio. We can calculate the likelihoods using Felsenstein's pruning algorithm (Box 8.1); then:

4. Find the product of the prior odds, proposal density ratio, and the likelihood ratio. In this case, both the prior odds and proposal density ratios are 1, so:

5. Draw a random number x from a uniform distribution between 0 and 1. If x < a, accept the proposed value of both parameters; otherwise reject, and retain the current value of the two parameters.

6. Repeat steps 2-5 a large number of times.

XXX Lizard example

Exploring Mk: the "total garbage" test

One problem that arises sometimes in maximum likelihood optimization happens when instead of a peak, the likelihood surface has a long flat "ridge" of equally likely parameter values. In the case of the Mk model, it is common to find that all values of q greater than a certain value have the same likelihood. This is because above a certain rate, evolution has been so rapid that all traces of the history of evolution of that character have been obliterated. After this point, character states of each lineage are random, and have no relationship to the shape of the phylogenetic tree. Our optimization techniques will not work in this case because there is no value of q that has a higher likelihood than other values. Once we get onto the ridge, all values of q have the same likelihood.

For Mk models, there is a simple test that allows us to recognize when the likelihood surface has a long ridge, and q values cannot be estimated. I like to call this test the "total garbage" test because it can tell you if your data are "garbage" with respect to historical inference – that is, your data have no information about historical patterns of trait change.

To carry out the total garbage test, imagine that you are just drawing trait values out of a hat. That is, each species has some probability p of having

character state 0, and some probability (1 - p) of having state 1 (one can also generalize this test to multi-state models). This model is easy to write down. For a tree of size n, the probability of drawing k species with state 0 is:

(eq. 8.2)

This equation gives the likelihood of the "total garbage" model for any value of p. Equation 8.2 is related to a binomial distribution (lacking only the factorial term). We know from probability theory that the ML estimate of p is k / n, with likelihood given by the above formula.

Now consider the likelihood surface of the Mk model. When Mk likelihood surfaces have long ridges, they are always for high values of q – and when the transition rate of character changes is high, this model converges to our "drawing from a hat" (or "garbage") model. The likelihood ridge lies at the value that is exactly taken from equation 8.2 above.

Thus, one can compare the likelihood of our Mk model to the total garbage model. If the maximum likelihood value of q has the same likelihood as our garbage model, then we know that we are on a ridge of the likelihood surface and q cannot be estimated. We also have no ability to make any statements about the past evolution of our character – in particular, we cannot estimate ancestral character state with any precision. By contrast, if the likelihood of the Mk model is greater than the total garbage model, then our data contains some historical information.

XXX Lizard example

Testing for differences in the forwards and backwards rate of character change

I have been referring to an example of flower evolution throughout this chapter, but we have not yet tested the hypothesis that I stated in the introduction: that transition rates from actinomorphy to zygomorphy are much higher than the reverse.

To do this, we can compare our one-rate Mk model with a two-rate model with differences in the rate of forwards and backwards transitions. This is a special case of the "all-rates different" model discussed in chapter two. Q matrices for these two models will be:

(eq. 8.3)

Notice that model one has one parameter, while the other has two. One can compare them using standard methods discussed in previous chapters – that is, a likelihood-ratio test, AIC, BIC, or other similar methods.

We can apply all of the above methods to analyze the evolution of limblessness in squamates. We can use the tree and character state data from Brandley et

al. (2008), which is plotted with ancestral state reconstructions as Figure 8.2.

If we fit an Mk model to these data assuming equal state frequencies at the root of the tree, we obtain a lnL of -80.5 and an estimate of the Q matrix as:

An extended-Mk model with different forward and backward rates gives a lnL of -79.4 and:

Note that the ARD model has a higher backwards than forwards rate; that is, we estimate a rate of gaining limbs that is higher than the rate of losing them! Is this statistically supported? We can compare the AIC scores of the two models. For the ER model, AICc = 163.0, while for the ARD model AICc = 162.8. The AICc score is higher for the unequal rates model, but only by about 0.2 – which is not definitive either way. So based on this analysis, we cannot rule out the possibility that forward and backward rates are equal.

A Bayesian analysis of the ARD model gives similar conclusions (Figure 8.3). We can see that the posterior distribution for the backwards rate (q21) is higher than the forwards rate (q12), but that the two distributions are broadly overlapping.

You might wonder about how we can reconcile these results, which suggest that squamates gain limbs at least as frequently as they lose them, with our biological intuition that limbs should be much more difficult to gain than they are to lose. But keep in mind that our comparative analysis is not using any information other than the states of extant species to reconstruct these rates. In particular, identifying irreversible evolution using comparative methods is a problem that is known to be quite difficult, and might require outside information in order to resolve conclusively. For example, if we had some information about the relative number of mutational steps required to gain and lose limbs, we could use an informative prior – which would, I suspect, suggest that limbs are more difficult to gain than they are to lose. Such a prior could dramatically alter the results presented in Figure 8.3. We will return to the problem of irreversible evolution later in the book (Chapter 13).

Chapter summary

In this chapter I describe how Felsenstein's pruning algorithm can be used to calculate the likelihoods of Mk and extended-Mk models on phylogenetic trees. I have also described both ML and Bayesian frameworks that can be used to test hypotheses about character evolution. This chapter also includes a description of the "total garbage" test, which will tell you if your data has information about evolutionary rates of a given character.

Analyzing our example of lizard limbs shows the power of this approach; we can estimate transition rates for this character over macroevolutionary time, and we can say with some certainty that transitions between limbed and limbless have been asymmetric. In the next chapter, we will build on the Mk model

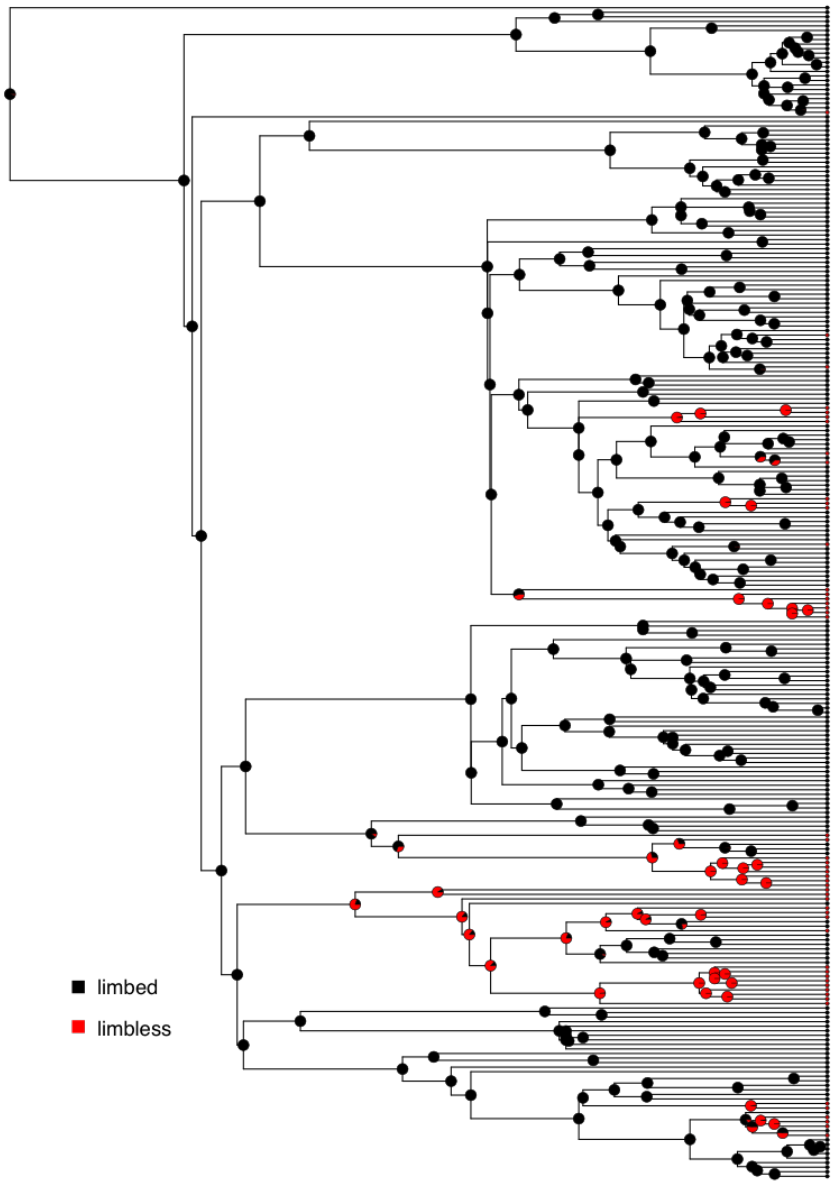

Figure 2: Figure 8.2. Reconstructed patterns of the evolution of limbs and limblessness across squamates. Tips show states of extant taxa (here, I classified species with neither fore- nor hindlimbs as limbless, which is conservative given the variation across this clade (see chapter 7). Pie charts on internal nodes show proportional marginal likelihoods for ancestral state reconstruction. Data from Brandley et al. 2008.

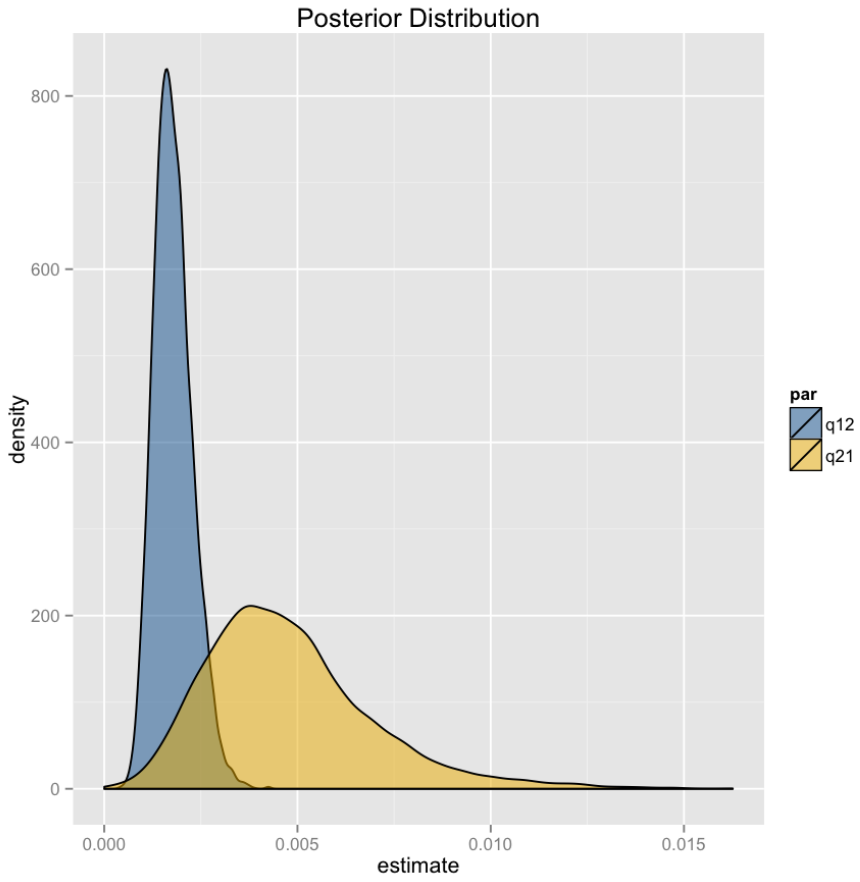

Figure 3: Figure 8.4. Bayesian posterior distibutions for the extended-Mk model applied to the evolution of limblessness in squamates.

and further develop our comparative toolkit for understanding the evolution of discrete characters.

Chapter 9: Beyond the Mk Model

Biological motivation: The Evolution of Frog Life History Strategies

Frog reproduction is one of the most bizarrely interesting topics in all of biology. Across the nearly 6,000 species of living frogs, one can observe a bewildering variety of reproductive strategies and modes. As children, we learn of the "classic" frog life history strategy: the female lays jellied eggs in water, which hatch into tadpoles, then later metamorphose into their adult form (e.g. Figure 9.1A). But this is really just the tip of the frog reproduction iceberg. There are foam-nesting frogs, which hang their eggs from leaves in foamy sacs over streams; when the eggs hatch, they drop into the water (Figure 9.1B). Male midwife toads carry fertilized eggs on their backs until they are ready to hatch, at which point they wade into water and their tadpoles wriggle free (Figure 9.1C). Perhaps most bizarre of all are the gastric-brooding frogs, now thought to be extinct. In this species, female frogs swallow their fertilized eggs, which hatch and undergo early development in their mother's stomach (Figure 9.1D). The young were then regurgitated to start their independent lives.

The great diversity of frog reproductive modes brings up several key questions that can potentially be addressed via comparative methods. How rapidly do these different types of reproductive modes evolve? Do they evolve more than once on the tree? Were "ancient" frogs more flexible in their reproductive mode than more recent species? Do some clades of frog show more flexibility in reproductive mode than others? To explore these questions, I will refer to a dataset of frog reproductive modes from Gomez-Mestre et al. (2012), specifically data classifying species as those that lay eggs in water, lay eggs on land without direct development (terrestrial), and species with direct development (Figure 9.2).

Many of the key questions stated above do not fall neatly into the Mk or extended-Mk framework presented in the previous characters. In this chapter, I will review approaches that elaborate on this framework and allow scientists to address a broader range of questions about the evolution of discrete traits.

Key Biological Questions

- Do transition rates among character states vary through time or across clades?
- Do discrete characters evolve in a correlated fashion?
- Does a model of underlying quantitative characters with thresholds explain discrete character data?

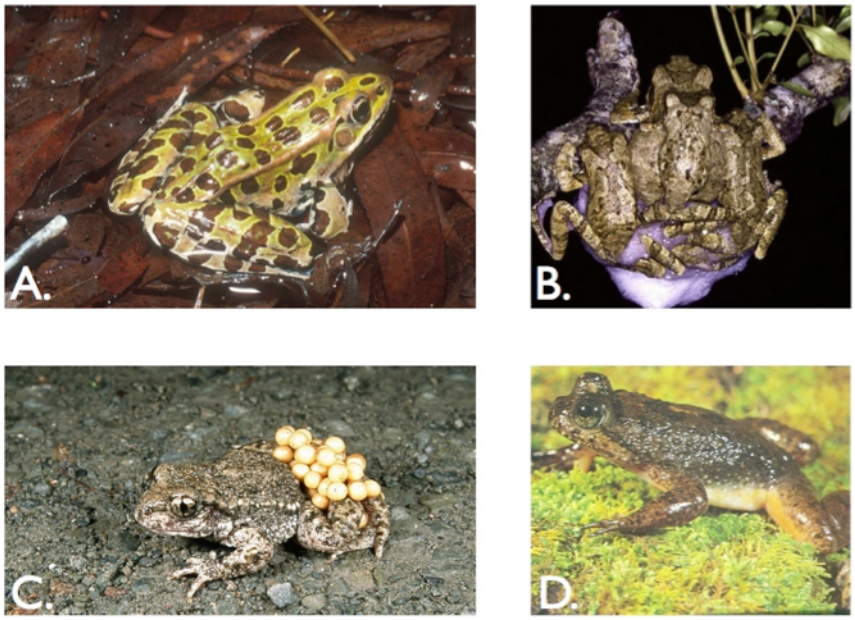

Figure 1: Figure 9.1. Examples of frog reproductive modes. (A) Leopard frogs lay jellied eggs in water, which hatch as tadpoles and metamorphose; (B) African foam-nesting frogs make nests that, supported by foam created during amplexus, hang from leaves and branches; (C) Male midwife toads carry fertilized eggs on their back; and (D) Female gastric-brooding frogs (now extinct) swallowed their fertilized eggs, which hatch and develop in the mother's stomach. Stolen without permission for now.

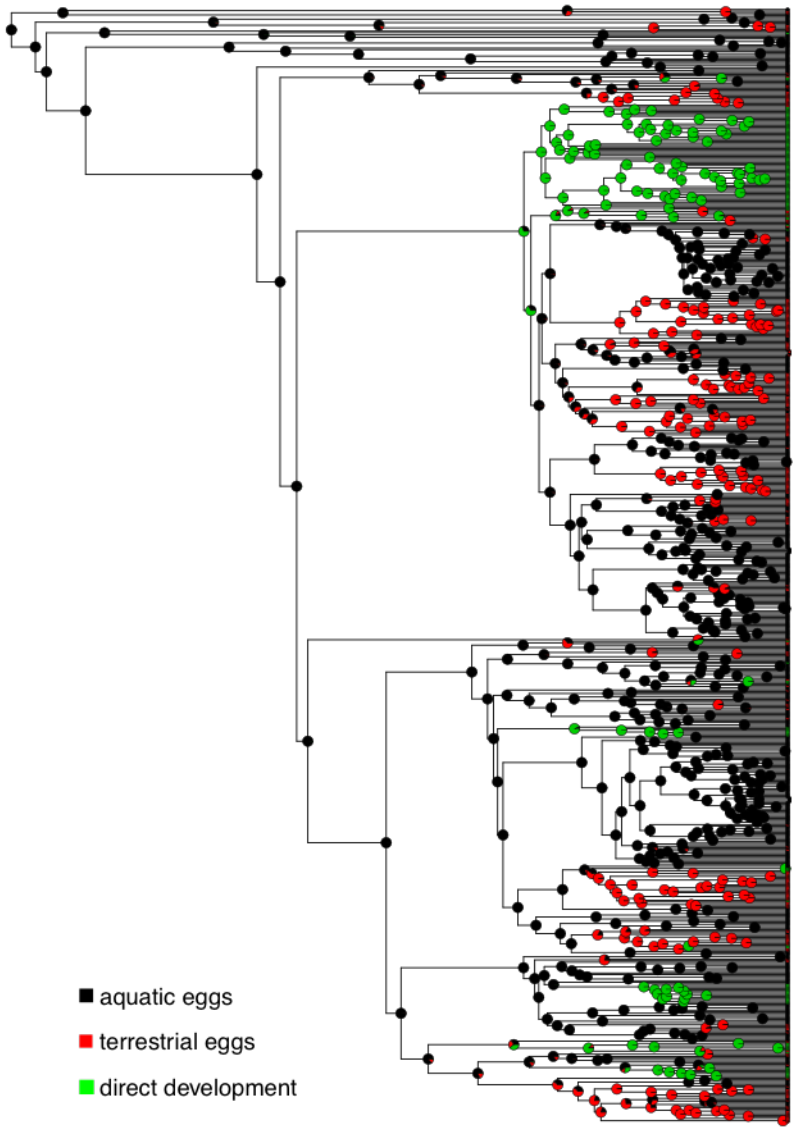

Figure 2: Figure 9.2. Ancestral state reconstruction of frog reproductive modes.

Beyond the Mk model

In Chapter 8, we considered the evolution of discrete characters on phylogenetic trees. These models fall under the general category of continuous-time Markov models, which consider a process that can occupy two or more states. Transitions occur between those states in continuous time. The Markov property means that, at some time t, what happens next in the model depends only on the current state of the process and not on anything that came before.

In evolutionary biology, the most detailed work on continuous time Markov models has focused on DNA or protein sequence data. An extremely large set of models are available for modeling and analyzing these molecular sequences. For example, there are dozens of evolutionary models that can be applied to genetic sequence data, ranging from the very simple (e.g. the Jukes-Cantor model [JC], a single-parameter model that assumes that all base frequencies are equal and transitions among them all occur at the same rate) to the complex (e.g. the general time-reversible model [GTR], which allows distinct transition rates between each pair of nucleotides and thus has 6 rate parameters and 3 free base frequency parameters [the fourth is set by the values of the other three]). One can also elaborate on these models by adding rate heterogeneity across sites (e.g. the gamma parameter, as in GTR + Γ), or other complications related to mechanisms of sequence evolution (for a review, see Lio and Goldman 1998).

However, there are two important differences between models of sequence evolution and models of character change on trees that make our task distinct from the task of modeling DNA or amino acid sequences. First, when analyzing molecular sequences, one typically has data for many thousands (or millions) of characters. Data sets for other characters – like the phenotypic characters of species – are typically much smaller (and harder to collect). Second, sequence analysis very often assumes that each character evolves independently from all other characters, but that all characters (or at least certain large subsets of those characters) evolve under a shared model. This means that, for example, the frequency of transitions between A and C at one location in a gene sequence contribute information about the same transition in a different location in the sequence. Unfortunately, when analyzing morphological character evolution, we are often interested in single characters, and the use of shared models across characters seems impossible to justify. There is usually no equivalence between different character states for different characters: an A is an A for sequences, but a "1" in a character matrix usually corresponds to the presence of two completely different characters. The consequence of this difference is reflected in the statistical property of multivariate data. For gene sequence problems, adding more data in the form of additional characters (sites) makes model-fitting easier, as each site adds information about the overall (shared) model across sites. With character data, additional characters do not make the problem any easier, because each character comes with its own model parameters. In fact, we will see that when considering character correlations using a generalized Mk model,

adding characters actually makes the problem more and more difficult. Perhaps these issues partially explain the slow pace of model development for fitting discrete characters to trees. There are a few potential solutions, such as threshold models (discussed below), but more work is desperately needed in this area.

In this chapter, we will first discuss extensions of Mk models that allow us to add complexity to this simple model. We also discuss threshold models, a relatively new approach in comparative methods that is distinct from Mk models and has some potential for future development.

Pagel's λ, δ, and κ (lambda, delta, and kappa)

The three Pagel models discussed in chapter 6 can also be applied to discrete characters. We do not create a phylogenetic variance-covariance matrix for species under an Mk model, so these three models can, in this case, only be interpreted in terms of transformations of the tree's branch lengths. However, the meaning of each parameter is the same as in the continuous case: λ scales the tree from its original form to a "star" phylogeny, and thus quantifies whether the data fits a tree-based model or one where all species are independent; δ captures changes in the rate of trait evolution through time; and κ scales branch lengths between their original values and one, and mimics a speciational model of evolution (but only if all species are sampled and there has been no extinction).

Just as with discrete characters, the three Pagel models can be evaluated in either an ML / AIC_c framework or using Bayesian analysis. One might expect these models to behave differently when applied to discrete rather than continuous characters, though. The main reason for this is that discrete characters, when they evolve rapidly, lose historical information surprisingly quickly. That means that models with high rates of character transitions will be quite similar to both models with low "phylogenetic signal" (i.e. $\lambda = 0$) and with rates that accelerate through time (i.e. $\delta > 0$).

We can apply these three models to data on frog reproductive modes. But first, we should try the Mk and extended-Mk models. Doing so, we find the following results:

Model	lnL	AIC_c	ΔAIC_c	AIC Weight
ER	-316.0	633.9	38.0	0.00
SYM	-296.6	599.2	3.2	0.17
ARD	-291.9	596.0	0.0	0.83

We can interpret this as strong evidence against the ER model, and weak support in favor of ARD over SYM. We can then try the three Pagel parameters. Since the support for SYM and ARD were similar, we can add extra parameters to each of them. Doing so, we obtain:

Model	Extra parameter	lnL	AIC_c	ΔAIC_c	AIC weight
ER		-316.0	633.9	38.0	0.00
ARD		-291.9	596.0	0	0.37
SYM	λ	-296.6	601.2	5.2	0.03
SYM	κ	-296.6	601.2	5.2	0.03
SYM	δ	-295.6	599.2	3.2	0.07
ARD	λ	-292.1	598.3	2.3	0.11
ARD	κ	-291.3	596.9	0.9	0.24
ARD	δ	-292.4	599.0	3.0	0.08

Notice that our results are somewhat ambiguous, with AIC weights spread fairly evenly across the three Pagel models. Interestingly, the overall lowest AIC score (and the most AIC weight, though only just more than 1/3 of the total) is on the ARD model with no additional Pagel parameters. I interpret this to mean that, for these data, the standard ARD model with no alterations is probably a reasonable fit to the data compared to the Pagel-style alternatives considered above.

Mk models where parameters vary across clades and/or through time

Another generalization of the Mk model we might imagine is a Mk model where rate parameters vary, either across clades or through time. There is some recent work along these lines, with two approaches that consider the possibility that rates of evolution for an Mk model vary on different branches of a phylogenetic tree (e.g. Marazzi et al. 2012, Beaulieu et al. 2013).

We can understand how these methods work in general terms by considering a simple case where the rate of character evolution is faster in one clade than in the rest of the tree. Ihis is the discrete-character version of Brownie, an approach for continuous characters that I discussed in chapter 6. The simplest way to implement a "discrete-trait Brownie" model is to incorporate variation across models into the pruning algorithm that is used to calculate the Mk model on a phylogenetic tree (see Fitzjohn 2013 for implementation). One can, for example, consider the most "Brownie-like" model possible under Mk: a model where the overall rate of evolution varies between clades in a phylogenetic tree. To do this, we can specify the background rate of evolution using some transition matrix $\mathbf{Q}$, and then assume that within our focal clade evolution can be modeled with some scalar value r, such that the new rate matrix is $r\mathbf{Q}$. Given $\mathbf{Q}$ and r, one can calculate the likelihood for this model using the pruning algorithm, modified in such a way that the appropriate transition matrix is used along each branch in the tree; one can then maximize the likelihood of the model for all parameters (those describing $\mathbf{Q}$, as well as r, which describes the relative rate of evolution in the focal clade compared to the background).

In even more general terms, we will consider the situation where we can describe the model of evolution using a set of $\mathbf{Q}$ matrices: $\mathbf{Q_1}$, $\mathbf{Q_2}$, ..., $\mathbf{Q_n}$, each of which can be assigned to a particular branch in a phylogenetic tree (or be assigned to branches depending on some other character that influences the rate of the focal character; Marazzi et al. 2012). The only limit here is that each $\mathbf{Q}$ matrix adds a new set of model parameters that must be estimated from the data, and it is easy to imagine this model becoming overparametized. If we imagine a model where every branch has its own $\mathbf{Q}$-matrix, then we are actually describing the "no common mechanism" model (Tuftey and Steel 1997; Steel and Penny 2000), which is statistically interchangeable with parsimony. It should also be possible to create a method that explores all models connecting simple Mk and the no common mechanism model using the machinery of reversible-jump MCMC, although I do not think such an approach has ever been implemented.

One can also imagine a situation where rate parameters in the $\mathbf{Q}$ matrix change through time. This might follow a constant pattern of increase or decrease through time, or might be related to some external driver like temperature. One can mimic models where rates change through time by changing the branch lengths of phylogenetic trees. If deep branches are lengthened relative to shallow branches, then we can fit a model where rates of evolution slow through time; conversely, lengthening shallow branches relative to deep ones creates a model where the overall rate of evolution accelerates through time (see Fitzjohn 2013).

More work could certainly be done in the area of time-varying rates of change. The most general approach is to write a set of differential equations that describe the changes in character state along single branches in the tree. Parameters in those equations can be made to vary, either through time or even in a way that is correlated with some external variable hypothesized to influence rates of change, like temperature or rainfall. Given such a model, the reverse-time approach of Maddison et al. (2007) can then be used to fit general time-varying (or even clade-varying) Mk models to data.

Threshold models

Recently, Joe Felsenstein (2005, 2012) introduced a model from quantitative genetics, the threshold model, to comparative method. Threshold models work by modeling a discrete character as underlain by some other, unobserved, continuous trait (called the liability). If the liability crosses a certain threshold value, then the discrete state changes. More specifically, we can consider a single trait, y, with two states, 0 and 1, which is in turn determined by some underlying continuous variable, x, called the liability. If x is greater than the threshold, t, then y is 1; otherwise, y is 0. Felsenstein (2005) assumes that x evolves under a Brownian motion model, although other models like OU are, in principle, possible.

We can find the likelihood to this model by considering the observations of

character states at the tips of the tree. We observe the state of each species, y_i. We do not know the liability values for these species. However, we treat these liabilities as unobserved and consider their distributions. Under a Brownian motion model, we know that the liabilities will follow a multivariate normal distribution (see chapter 3). We can calculate the probability of observing the data (y_i) by finding the integral of the distributions of liabilities on the side of the threshold that matches the data. So if the distribution of the liability for species i is $p_i(x)$, then:

(eq 9.1)

$$p(y_i = 0) = \int_{-\infty}^{t} p_i(x)dx$$

and

$$p(y_i = 1) = \int_{t}^{\infty} p_i(x)dx$$

(see Figure 9.3 for an illustration of this calculation).

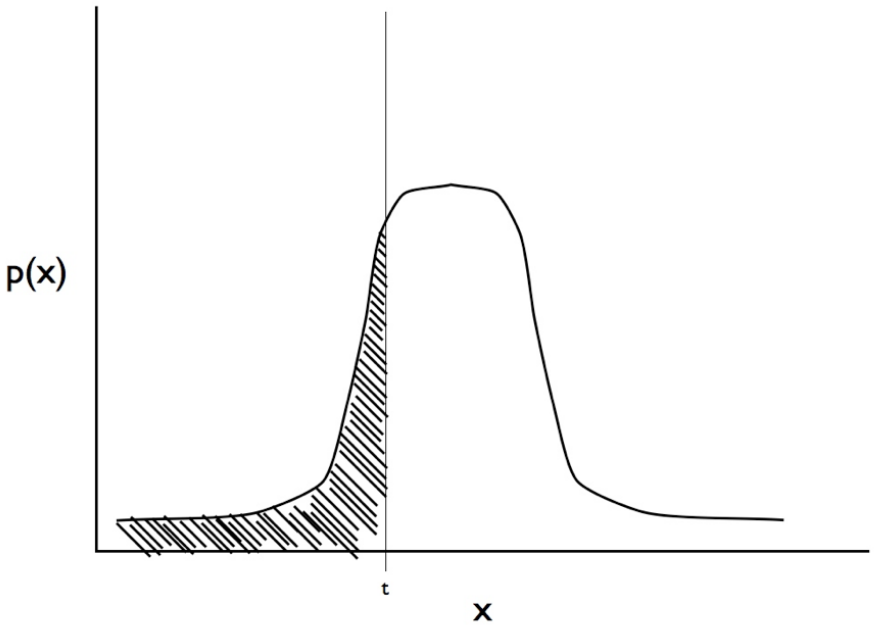

Figure 3: Figure 9.3. Illustration of the integral in equation 9.1.

One can fit this model using either an expectation-maximization (EM) algorithm (Felsenstein 2005, 2011) or a Bayesian MCMC (Revell 2012).

The threshold model differs in some key ways from standard Mk-type models. First of all, threshold characters evolve differently than non-threshold characters because of their underlying liability. In particular, the effective rate of change of the discrete character depends on the amount of time that a lineage has been in that character state. Characters that have just changed (say, from 0 to 1) are likely to change back (from 1 to 0), since the liability is likely to be near the threshold. By contrast, characters that have been in one state or the other for a long time are more unlikely to change (since the liability is likely very far from the threshold). This difference matches biological intuition for some characters, where millions of years in one state means that change to a different state might be unlikely. This behavior of the threshold model can potentially account for variation in transition rates across clades without adding additional model parameters. Second, the threshold model scales to cover more than one character more readily than Mk models. Finally, in a threshold framework, it is straightforward to extend the model to include a mixture of both discrete and continuous characters – basically, one assumes that the continuous characters are like "observed liabilities," and can be modeled together with the discrete characters.

Modeling more than one discrete character at a time

It is extremely common to have datasets with more than one discrete character – in fact, one could argue that multivariate discrete datasets are the cornerstone of systematics. Nowadays, the most common multivariate discrete datasets are composed of genetic/genomic data. However, the foundations of modern phylogenetic comparative biology were laid out by Hennig and the other early cladists, who worked out methods for using discrete character data to obtain phylogenetic trees that show the evolutionary history of clades.

Almost all phylogenetic reconstruction methods that use discrete characters as data make a key assumption: that each of these characters evolves independently from one another. Mathematically, one calculates the likelihood for each single character, then multiplies this likelihood (or, equivalently, adds the log-likelihood) across all characters to obtain the likelihood of the data.

The assumption of character independence is clearly not true in general. In the case of morphological characters, structures often interact with one another to determine the fitness of an individual, and it seems very likely that those structures are not independent. In fact, some times we are specifically interested in whether or not particular sets of characters evolve independently or not. Methods that assume character independence a priori are not useful for that sort of framework.

Felsenstein (1985) made a huge impact on the field of evolutionary biology with

a statistical argument about species: species can not be considered independent data points because they share an evolutionary history. Species that are most closely related to one another will covary, simply due to that shared history. Nowadays, one cannot publish a paper in comparative biology without accounting directly for the non-independence of species that evolve on a tree. However, it is still very common to ignore the non-independence of characters, even when they occur together in the same organism! Surely the shared developmental history of two characters within one body commonly leads to correlations across these characters.

Testing for correlated evolution of different characters

Hypotheses in evolutionary biology often relate to whether two (or more) traits evolve in a correlated manner. The situation is similar to what I discussed for correlations of continuous characters in chapter 5. One can have a standard correlation between two discrete traits if knowing the state of one trait allows you to predict the state of the other. However, in evolution, these correlations will arise due to the shared patterns of relatedness across species. We are typically more interested in evolutionary correlations. With discrete traits, we can state evolutionary correlations in a more specific way: two discrete traits share an evolutionary correlation if the state of one character affects the transition rates of a second.

Imagine that we are considering the evolution of two traits, trait 1 and trait 2, on a phylogenetic tree. Both traits have two possible character states, one and zero. We can show these two traits visually as Figure 9.4.

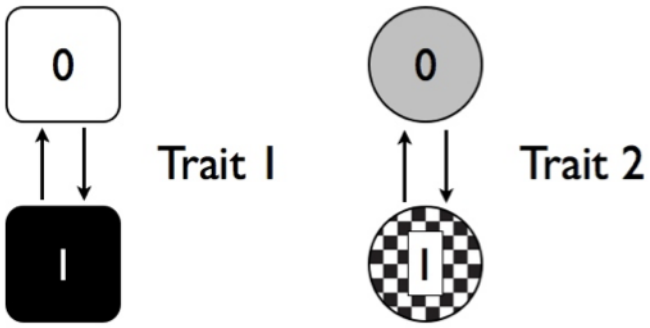

Figure 4: Figure 9.4. Two discrete character traits, each with two states (labeled 0 and 1).

In the figure, each trait has two possible transition rates, from 0 to 1 and from 1 to 0. For now, let's assume that backwards and forward rates are equal. Any species can have one of four possible combinations of the two traits (00, 01, 10,

or 11). We can draw the transitions among these four combinations as Figure 9.5:

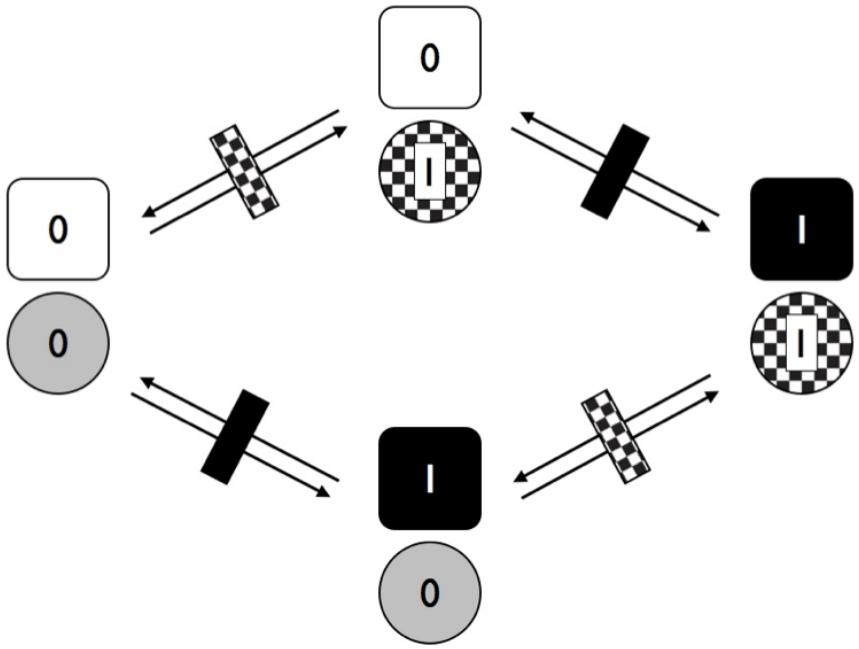

Figure 5: Figure 9.5. Transitions among states for two traits with two character states each where characters evolve independently of one another.

In Figure 9.6, I have marked the distinct rates with different rectangles – black represents changes in trait 1, while checkered is changes in trait 2. Notice that, in this figure, we are assuming that the two traits are independent. That is, in this model the transition rates of trait one do not depend on the state of trait 2, and vice-versa. What would happen to our model if we allow the traits to evolve in a dependent manner?

Notice that in Figure 9.6, we have four different transition rates. Consider first the solid rectangles. The grey rectangle represents the transition rate for trait 1 when trait 2 has state 0, while the black rectangle represents the transition rate for trait 1 when trait 2 has state 1. If these two rates are different, then the traits are dependent on each other – that is, the rate of evolution of trait 1 depends on the character state of trait 2.

These two models have different numbers of parameters, but are relatively easy to fit using the maximum-likelihood approach outlined in this chapter. The key is to write down the transition matrix ($\mathbf{Q}$) for each model. For example, a transition matrix for model in figure 9.4 is:

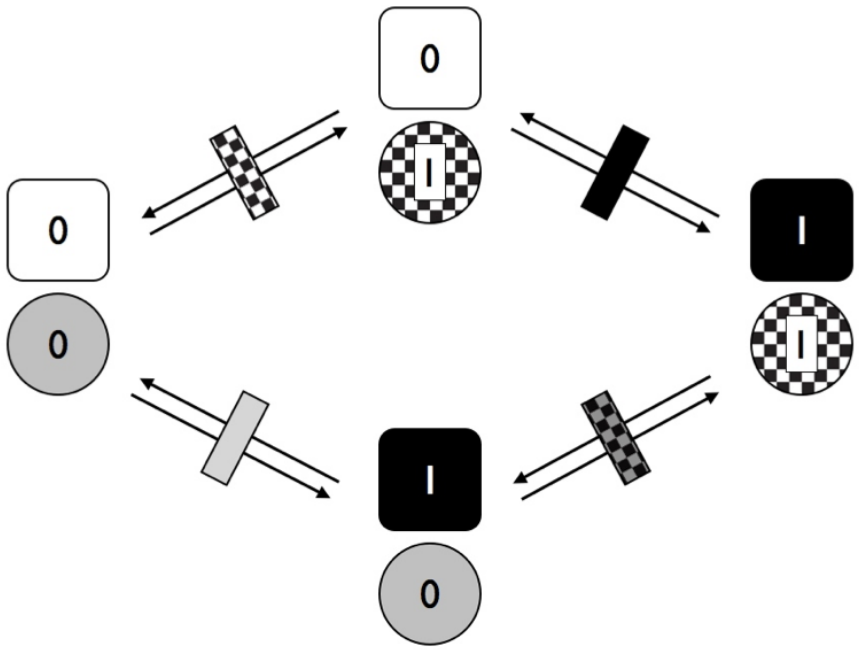

Figure 6: Figure 9.6. Transitions among states for two traits with two character states each where characters evolve at rates that depend on the character state of the other trait.

(eq. 9.2)

$$\mathbf{Q} = \begin{bmatrix} -q_1 - q_2 & q_1 & q_2 & 0 \\ q_1 & -q_1 - q_2 & 0 & q_2 \\ q_2 & 0 & -q_1 - q_2 & q_1 \\ 0 & q_2 & q_1 & -q_1 - q_2 \end{bmatrix}$$

In the matrix above, each row and column corresponds to a particular combination of states for character 1 and 2: (0,0), (0,1), (1,0), and (1,1). Note that some possible transitions in this model have rate 0, meaning they do not occur. These are transitions that would require both characters to change exactly simultaneously (e.g. (0,0) to (1,1) – a possibility that is excluded from this model.

Similarly, we can write a transition matrix for the model in figure 9.5:

(eq. 9.3)

$$\mathbf{Q} = \begin{bmatrix} -q_1 - q_2 & q_1 & q_2 & 0 \\ q_1 & -q_1 - q_3 & 0 & q_3 \\ q_2 & 0 & -q_2 - q_4 & q_4 \\ 0 & q_3 & q_4 & -q_3 - q_4 \end{bmatrix}$$

Notice that the simple, 2-parameter independent evolution model is a special case of the more complex, 4-parameter dependent model. Because of this, we can compare the two with a likelihood ratio test. Alternatively, AIC or Bayes factors can be used. If we find support for the 4-parameter model, we can conclude that the two characters have an evolutionary correlation.

It is worth noting that there are other models that one can fit for the evolution of two binary traits that I did not discuss above. For example, one can model the situation where the two traits each have different forwards and backwards rates, but are evolving independently. This is a four-parameter model. Additionally, one can allow both forward and backward rates to differ and to depend on the character state of the other trait: an eight-parameter model. All of these models – and others not described here – can be compared using AIC, BIC, or Bayes Factors. Pagel and Meade (2006) describe a particularly innovative and synthetic method to test hypotheses about correlated evolution of discrete characters in a Bayesian framework using reversible-jump MCMC.

One can also test for correlations among discrete characters using threshold models. Here, one assumes that the liabilities for the two characters evolve in a correlated fashion. More specifically, we can model liabilities for the two threshold characters using a bivariate Brownian motion model, with some evolutionary covariance σ^2_{12} between the two liabilities. We can then use either ML or Bayesian methods to determine if the evolutionary covariance between the two characters is non-zero (following the methods described in chapter 5, but using likelihoods based on discrete characters as described above.

Chapter 10: Introduction to birth-death models

Introduction: Plant diversity imbalance

The diversity of flowering plants (the angiosperms) dwarfs the diversity of their closest evolutionary relatives (Figure 10.1). There are more than 260,000 species of angiosperms (that we know; more are added every day). The clade originated some 130 million years ago, so all of these species have formed since then. One can contrast the diversity of angiosperms with the diversity of other groups that originated at around the same time. For example, gymnosperms, which are as old as angiosperms, include only around 1000 species, and may even represent more than one clade. The diversity of angiosperms also dwarfs the diversity of familiar vertebrate groups of similar age (e.g. squamates, snakes and lizards, which diverged from their sister taxon, the tuatara, some 250 mya or more, include fewer than 8000 species).

The evolutionary rise of angiosperm diversity puzzled Darwin over his career, and the issues surrounding angiosperm diversification are often referred to as "Darwin's abominable mystery" in the scientific literature (e.g. Davies et al. 2004). The main mystery is the tremendous variation in numbers of species across plant clades (see Figure 10.1). This variation even applies within angiosperms, where some clades are much more diverse than others.

At a global scale, the number of species in a clade can change only via two processes: speciation and extinction. This means that we must look to speciation and extinction rates – and how they vary through time and across clades – to explain phenomena like the extraordinary diversity of Angiosperms. It is to this topic that we turn in the next few chapters. Since Darwin's time, we have learned an extraordinary amount about the evolutionary processes that led to the diversity of angiosperms that we see today. These data provide an incredible window into the causes and effects of speciation and extinction over macroevolutionary time scales.

Comparative methods can be applied to understand patterns of species richness, both across clades and through time, by estimating speciation and extinction rates. In this chapter, I will introduce birth-death models, by far the most common model for understanding diversification in a comparative framework. I will discuss the mathematics of birth-death models and how these models relate to the shapes of phylogenetic trees. I will describe how to simulate phylogenetic trees under a birth-death model. Finally, I will discuss tree balance and lineage-through-time plots, two common ways to measure the shapes of phylogenetic trees.

Key Questions

- How can we use birth-death models to model diversification?

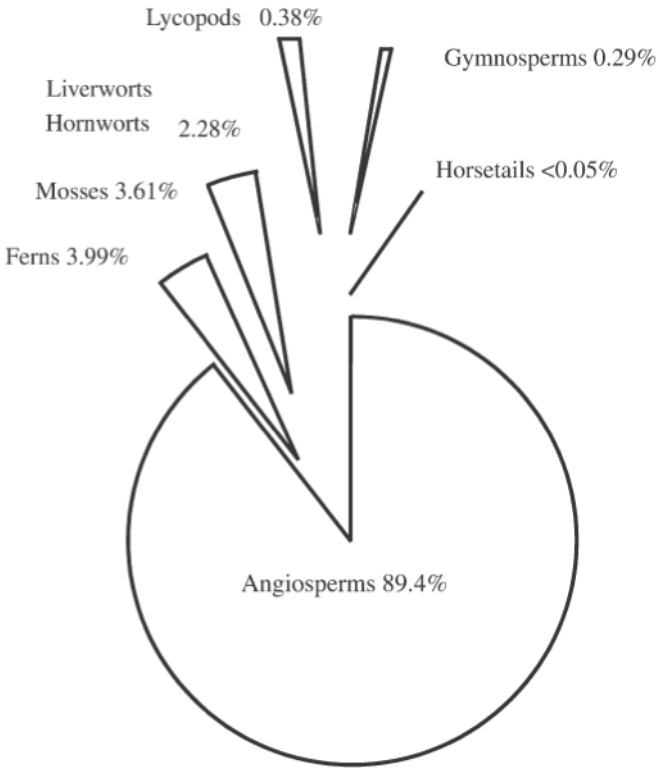

Figure 1: Figure 10.1. Diversity of major groups of embryophytes (land plants). Angiosperms, including some 250,000 species, comprise more than 90% of species of land plants. Figure taken from Crepet and Niklas 2009

- How do we simulate phylogenetic trees under a birth-death model?
- How do birth-death models relate to the topology and balance of phylogenetic trees?
- What is the shape of lineage-through-time plots under birth-death models?

The birth-death model

A birth-death model is a continuous-time Markov process that is typically used to study how the number of individuals in a population change through time. (For macroevolution, these "individuals" are usually species). In a birth-death model, two things can occur: births, where the number of individuals increases by one; and deaths, where the number of individuals decreases by one. We assume that no more than one new individual can form during any one event. In phylogenetic terms, that means that birth-death trees cannot have "hard polytomies."

In macroevolution, we apply the birth-death model to species, and typically consider a model where each lineage has a constant probability of either giving birth (speciating) or dying (going extinct). We denote the per-lineage birth rate as and the per-lineage death rate as . For now we consider these rates to be constant, but we will relax that assumption later in the book.

We can understand the behavior of birth-death models if we consider the waiting time between successive speciation and extinction events in the tree. Imagine that we are considering a single lineage that exists at time t0. We can think about the waiting time to the next event, which will either be a speciation event splitting that lineage into two (Figure 10.2A) or an extinction event marking the end of that lineage (Figure 10.2B). Under a birth-death model, both of these events follow a Poisson process, so that the expected waiting time to an event follows an exponential distribution (Figure 10.2C). The expected waiting time to the next speciation event is exponential with parameter , and the expected waiting time to the next extinction event exponential with parameter . Of course, only one of these can be the next event. The expected waiting time to the next event (of any sort) is exponential with parameter + , and the probability that that event is speciation is / (+), extinction / (+).

When we have more than one lineage "alive" in the tree at any time point, then the waiting time to the next event changes, although its distribution is still exponential. In general, if there are N(t) lineages alive at time t, then the waiting time to the next event follows an exponential distribution with parameter N (+), with the probability that that event is speciation or extinction the same as given above. Using this approach, we can grow phylogenetic trees of any size (Figure 10.2D).

We can derive some important properties of the birth-death process on trees. To do so, it is useful to define two additional parameters, the net diversification rate (r) and the relative extinction rate ( ):

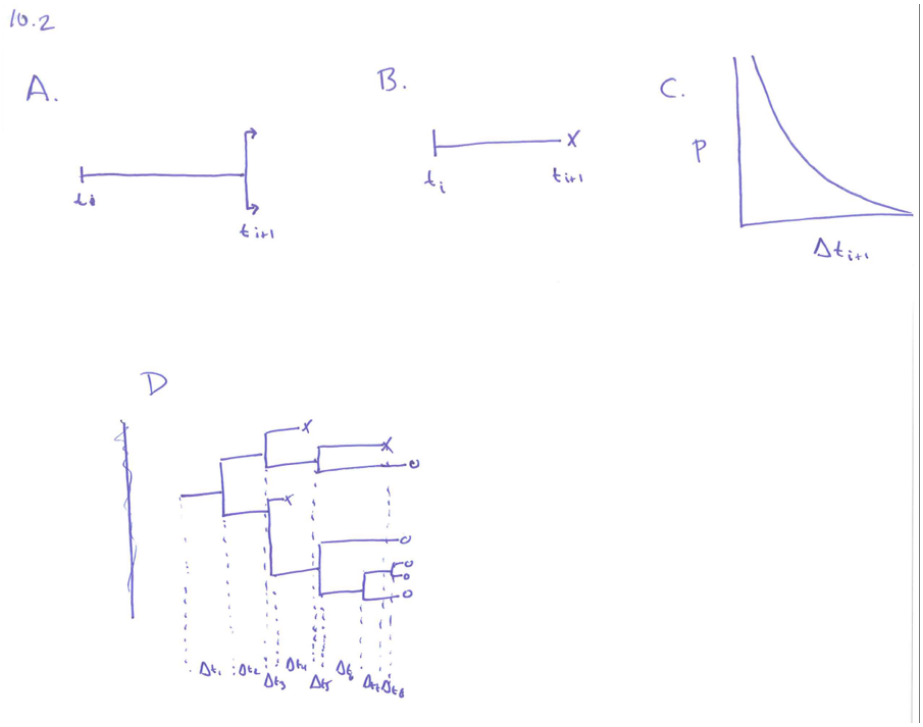

Figure 2: Figure 10.2. Illustration of the basic properties of birth-death models. A. Waiting time to a speciation event; B. Waiting time to an extinction event; C. Exponential distribution of waiting times until the next event; D. A birth-death tree with waiting times.

(10.1) r= - = /

These two parameters simplify some of the equations below, and are also commonly encountered in the literature.

To derive some general properties of the birth-death model, we first consider the process over a small interval of time, Δt. We assume that this interval is so short that it contains at most a single event, either speciation or extinction (the interval might also contain no events at all). The probability of speciation and extinction over the time interval can be expressed as:

(10.2)

We now consider the total number of living species at some time t, and write this as N(t). It is useful to think about the expected value of N(t) under a birth-death model [we consider the full distribution of N(t) below]. The expected value of N(t) after a small time interval Δt is:

(10.3) N(t+Δt)=N(t)+N(t) Δt-N(t) Δt

We can convert this to a differential equation by subtracting N(t) from both sides, then dividing by Δt and taking the limit as Δt becomes very small:

(10.4) dN/dt=N(-)

We can solve this differential equation if we set a boundary condition that N(0)=a; that is, at time 0, we start out with a lineages. We then obtain:

(10.5) $N(t)=ae^{(-)t=ae}rt$

This deterministic equation gives us the expected value for the number of species through time under a birth-death model. Notice that the number of species grows exponentially through time as long as > , e.g. r>0, and decays otherwise (Figure 10.3).

We are also interested in the stochastic behavior of the model – that is, how much should we expect N(t) to vary from one replicate to the next? We can calculate the full probability distribution for N(t), which we write as p_n (t)=Pr[N(t)=n] for all n 0, to completely describe the birth-death model's behavior. To derive this probability distribution, we can start with a set of equations, one for each value of n, which we will denote as p_0 (t) (there are an infinite set of such equations, from p_0to p_∞). We can then write a set of difference equations that describe the different ways that one can reach any state over some small time interval Δt. We again assume that Δt is sufficiently small that at most one event (a birth or a death) can occur. As an example, for n = 0, we can either be at n = 0 at the beginning of the time interval, or be at n = 1 and have the last surviving lineage go extinct. We write this as:

(10.6) p_0 (t+Δt)=p_1 (t) Δt+p_0 (t)

For any n 1, we can reach the state of n lineages in three ways: from a birth (from n – 1 to n), a death (from n + 1 to n), or neither (from n to n). This is

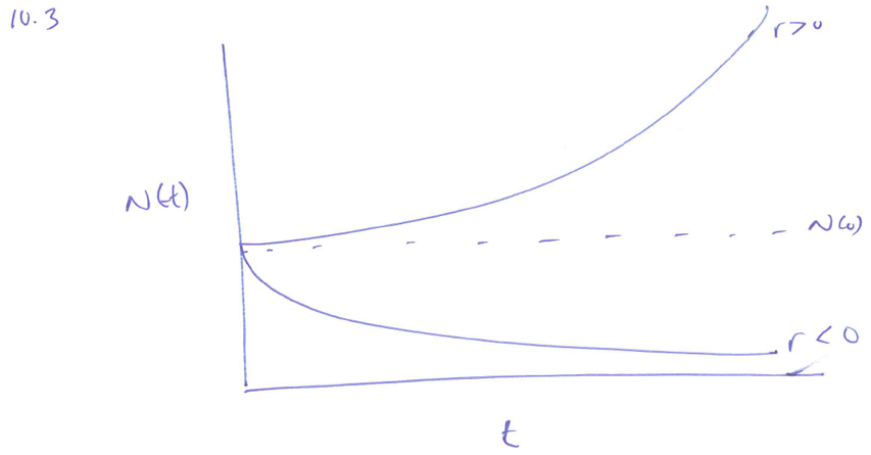

Figure 3: Figure 10.3. Expected number of species under a birth-death model with r= - > 0 and r < 0.

written as:

(10.7)

p_n (t+Δt)=p_(n-1) (t)(n-1) Δt+p_(n+1) (t)(n+1) Δt+p_n (t)(1-n(+))Δt

We can convert this set of difference equations to differential equations by subtracting p_n (t) from both sides, then dividing by Δt and taking the limit as Δt becomes very small. So, when n = 0, we use 10.6 to obtain:

(10.8) (dp_0)/dt= p_1

and, from 10.7, for all n 1:

(10.9) (dp_n)/dt=(n-1) p_(n-1)+(n+1) p_(n+1)-n(+) p_n

We can then solve this set of differentia l equations to obtain the probability distribution of pn(t). Using the same boundary condition, N(0)=a, we have p_0 (t)=1 for n = a and 0 otherwise. Then, we can find the solution to the differential equations 10.8 and 10.9. The derivation of the solution to this set of differential equations is beyond the scope of this book (but see Kot 2001 for a nice explanation of the mathematics). A solution was first obtained by Bailey (1964), but I will use the simpler equivalent form from Foote et al. (1999). For p_0 (t) – that is, the probability that the entire lineage has gone extinct at time t – we have:

(10.10) p_0 (t)= ^a

And for all n 1:

(10.11)

Where:

(10.12) $= (e^{rt-1)/(e}rt-) = (e^{rt-1)/(e}rt-)$

We can first write down the probability that this lineage has gone extinct; this means it has left no descendants at time t:

(10.13) _1 (t)=Pr[N(t)=0]= $(e^{rt-1)/(e}rt-)$

Note that when a = 1 – that is, when we start with a single lineage - equations 10.10 and 10.11 simplify to (Raup 1985):

(10.14) p_0 (t)=

And for all n 1:

(10.15) p_n (t)=(1-)(1-) ^(i-1)

In all cases the expected number of lineages in the tree is exactly as stated above in equation (10.5), but now we have the full probability distribution of the number of lineages given a, t, , and . A few plots capture the general shape of this distribution (Figure 10.4).

There are quite a few comparative methods that use clade species richness and age along with the distribution defined in 10.14 and 10.15 to make inferences about clade diversification rates (see chapter 11).

Birth-death models and phylogenetic trees

The above discussion considered the number of lineages under a birth-death model, but not their phylogenetic relationships. However, just by keeping track of the parent-offspring relationships among lineages, we can consider birth-death models that result in phylogenetic trees (e.g. Figure 10.2D).

The main complication in phylogenetic studies of birth-death models is that we get a "censored" view of the process, in that we only observe lineages that survive to the present day. In the above example, if the true phylogenetic tree were the one plotted in 10.5A, we would only have a chance to observe the phylogenetic tree in figure 10.5B – and even then only if we sampled all of the species and reconstructed the tree with perfect accuracy! A partially sampled tree with only extant species can be seen in Figure 10.5C. I will cover the relationship between birth-death models and the branch lengths of phylogenetic trees in much more detail in the next chapter.

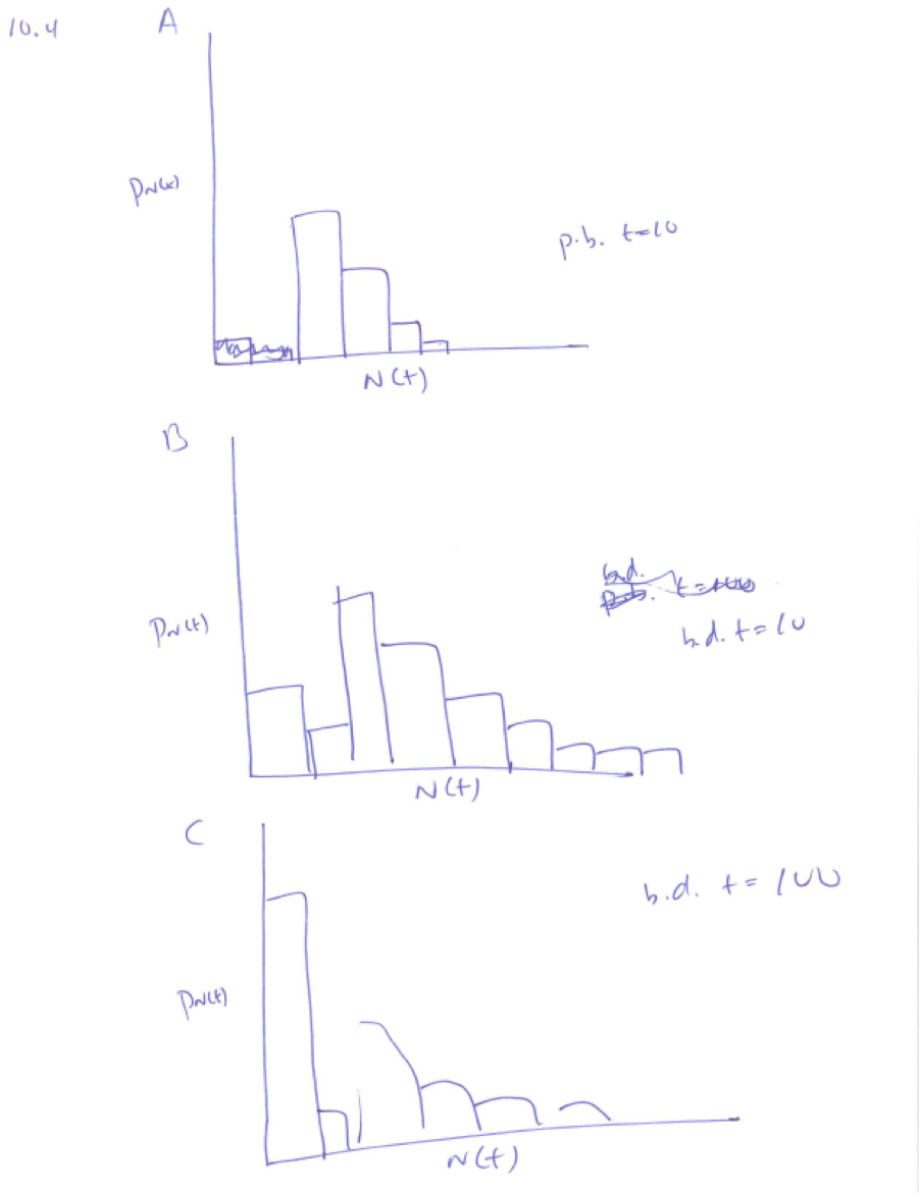

Figure 4: Figure 10.4. Probability distributions of N(t) under A. pure birth, B. birth death after a short time, and C. birth-death after a long time.

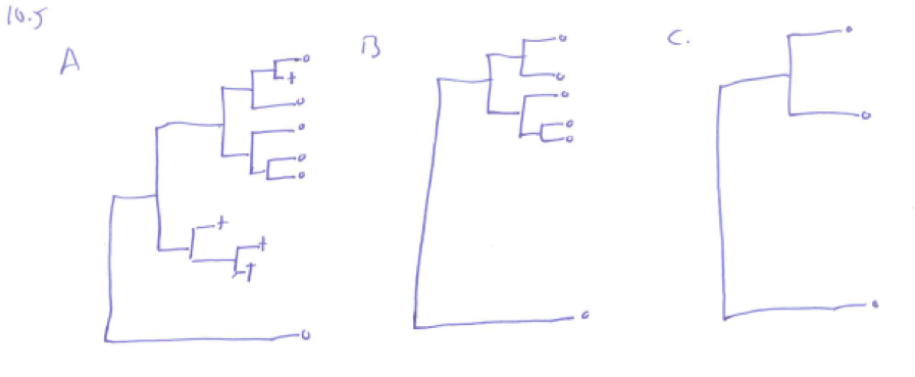

Figure 5: Figure 10.5. A. A birth-death tree including all extinct and extant species; B. A birth-death tree including only extant species; and C. A partially sampled birth-death tree including only some extant species.

Simulating birth-death trees

We can use the statistical properties of birth-death models to simulate phylogenetic trees through time. We could begin with a single lineage at time 0. However, phylogenetic tree often start with the first speciation event in the clade, so one can also begin the simulation with two lineages at time 0 (this distinction relates to our earlier discussion of crown versus stem ages; see also chapter 11).

To simulate our tree, we need to draw waiting times between speciation and extinction events, connect new lineages to the tree, and prune lineages when they go extinct. We also need a stopping criterion, which can have to do with a particular number of taxa or a fixed time interval. We will consider the latter, and leave growing trees to a fixed number of taxa as an exercise for the reader. Our simulation algorithm is as follows. I assume that we have a certain number of "living" lineages in our tree (1 or 2 initially), a current time (tc = 0 initially), and a stopping time tstop.

1. Draw a waiting time ti to the next speciation or extinction event. Waiting times are drawn from an exponential distribution with rate parameter NA * (+) where NA is the current number of living lineages in the tree.
2. Check to see if the simulation ends before the next event. That is, if tc + ti > tstop, end the simulation.
3. Decide whether the next event is a speciation event [with probability / (+)] or an extinction event [with probability / (+)]. This can be done by drawing a uniform random number ui from the interval [0,1] and assigning speciation to the event if ui < / (+) and extinction otherwise.
4. If (3) is a speciation event, then choose a random living lineage in the tree.

Attach a new branch to the tree at this point, and add one new living lineage to the simulation. Return to step 1.

5. If (3) is an extinction event, choose a random living lineage in the tree. That lineage is now dead. As long as there is still at least one living lineage in the tree, return to (1); otherwise, your whole clade has gone extinct, and you can stop the simulation.

This procedure returns a phylogenetic tree that includes both living and dead lineages. One can prune out any extinct taxa to return a birth-death tree of survivors, which is more in line with what we typically study using extant species. It is also worth noting that entire clades can – and often do – go extinct under this protocol before one reaches time tstop.

We can think about phylogenetic predictions of birth-death models in two ways: by considering tree topology, and by considering tree branch lengths. I will consider each of these two aspects of trees below.

Tree topology, tree shape, and tree balance under a birth-death model

Tree topology summarizes the patterns of evolutionary relatedness among a group of species independent of the branch lengths of a phylogenetic tree. Two different trees have the same topology if they define the exact same set of clades. This is important because sometimes two trees can look very different and yet still have the same topology (e.g. Figure 10.6 A, B, and C).

Tree topology ignores both branch lengths and tree tip labels. For example, the two trees in figure 10.6 A and D have the same tree topology even though they share no tips in common. What they do share is that their nodes have the same patterns in terms of the number of descendants on each "side" of the bifurcation. By contrast, the phylogenetic tree in 10.6 E has a different topology. (Note that what I am calling tree topology is sometimes referred to as "unlabeled" tree topology; e.g. Felsenstein 2004).

Finally, tree balance is a way of expressing differences in the number of descendants between pairs of sister lineages at different points in a phylogenetic tree. For example, consider the phylogenetic tree depicted in figure 10.6D. The deepest split in that tree separates a clade with five species (lizard, snake, turtle, frog, salamander) from a clade with a single species (trout), and so that node in the tree is unbalanced with a (5, 1) pattern. By contrast, the deepest split in 10.6E separates two clades of equal size. In that tree, the deepest node is balanced with a (3, 3) pattern. A number of approaches in macroevolution use balance at nodes and across whole trees to try to capture important evolutionary patterns.

We can start to understand these approaches by considering the balance of a single node n in a phylogenetic tree. There are two clades descended from this node; let's call them a and b. We assume that the total number of species

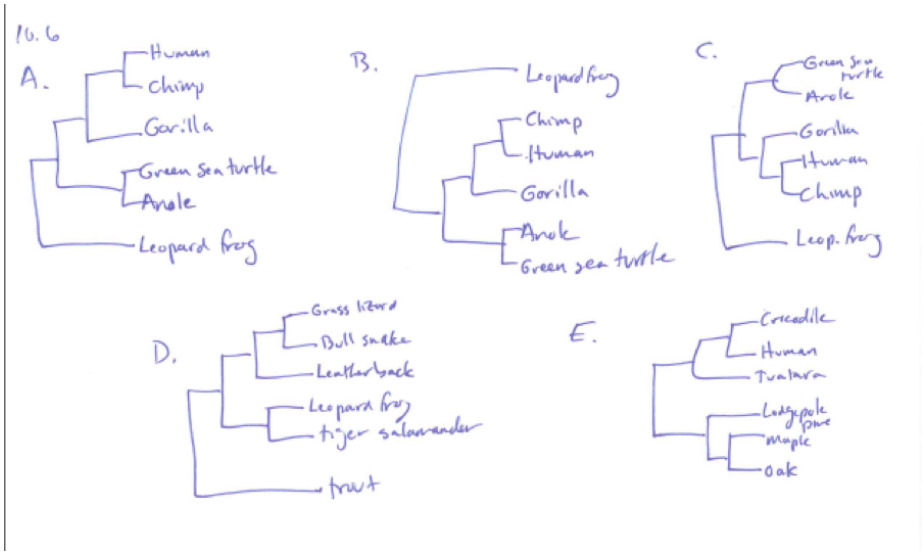

Figure 6: Figure 10.6. Several phylogenetic trees showing different ways to plot the same tree shape.

descended from the node Nn = Na + Nb is constant and that neither Na nor Nb is zero. An important result, first discussed by Farris (1976) for a pure-birth model, is that all possible numerical divisions of Nn into Na + Nb are equally probable. For example, if Nn = 10, then all possible divisions: 1+9, 2+8, 3+7, 4+6, 5+5, 6+4, 7+3, 8+2, and 9+1 are all equally probable, so that each will be predicted to occur with a probability 1/9. Formally

(eq. 10.2) p(N_a N_n)=1/(N_n-1)

(Note that there is a subtle difference between equation 10.2 above and some equations in the literature, e.g. Slowinsky and Guyer 1993. Ths difference has to do with whether we label the two descendent clades, a and b, or not; if the clades are unlabeled, then there is no difference between 4+6 and 6+4, so that the probability that the largest clade, whichever it might be, has 6 species is twice what is given by my equation). Equation 10.2 applies even if there is extinction, as long as both sister clades have the same speciation and extinction rates (Slowinsky and Guyer 1989). This equation has been used to compare diversification rates between sister clades, either for a single pair or across multiple pairs (see Chapter 11).

Tree balance statistics provide a way of comparing numbers of taxa across all of the nodes in a phylogenetic tree simultaneously. There are a surprisingly large number of tree balance statistics, but all rely on summarizing information about the balance of each node across a whole tree. Colless' index Ic is one of the simplest – and, perhaps, most commonly used – indices of tree balance. Ic is

the sum of the difference in the number of tips subtended on each side of every node in the tree, standardized by the maximum that such a sum can achieve:

(eq. 10.3)

If the tree is perfectly balanced (only possible when N is some power of 2, e.g. 2, 4, 8, 16, etc.), then IC = 0 (Figure 10.7C). By contrast, if the tree is completely pectinate, which means that each split in the tree contrasts a clade with 1 species with the rest of the species in the clade, then IC = 1 (Figure 10.7A). Most phylogenetic trees have values of IC between 0 and 1 (Figure 10.7B).

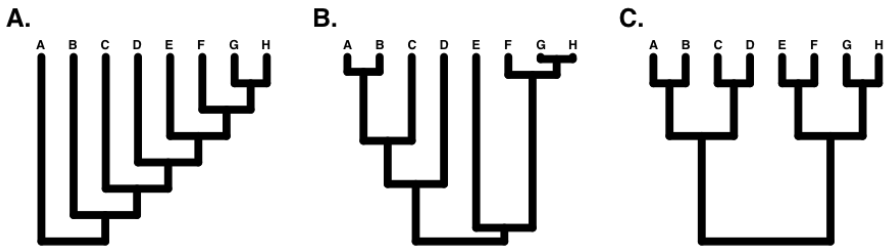

Figure 7: Figure 10.7. A. a pectinate tree; B. a random tree; C. A balanced tree.

There are a number of other indices of phylogenetic tree balance (reviewed in Mooers and Heard xxx; and others). All of these indices are used in a similar way: one can then compare the value of the tree index to what one might expect under a particular model of diversification, typically birth-death. In fact, since these indices focus on tree topology and ignore branch lengths, one can actually consider their general behavior under a set of equal-rates Markov (ERM) models. This set includes any model where birth and death rates are equal across all lineages in a phylogenetic tree at a particular time. ERM models include birth-death models as described above, but also encompass models where birth and/or death rates change through time.

Lineage-through-time plots

The other main way to quantify phylogenetic tree shape is by making lineage-through-time plots. These plots have time along the x axis (from the root of the tree to the present day), and the reconstructed number of lineages on the y-axis (Figure 10.8). Since we are usually considering birth-death models, where the number of lineages is expected to grow (or shrink) exponentially through time, then it is typical practice to log-transform the y-axis.

Lineage-through-time plots are effective ways to visualize patterns of lineage diversification through time. Under a pure-birth model, LTT plots follow a

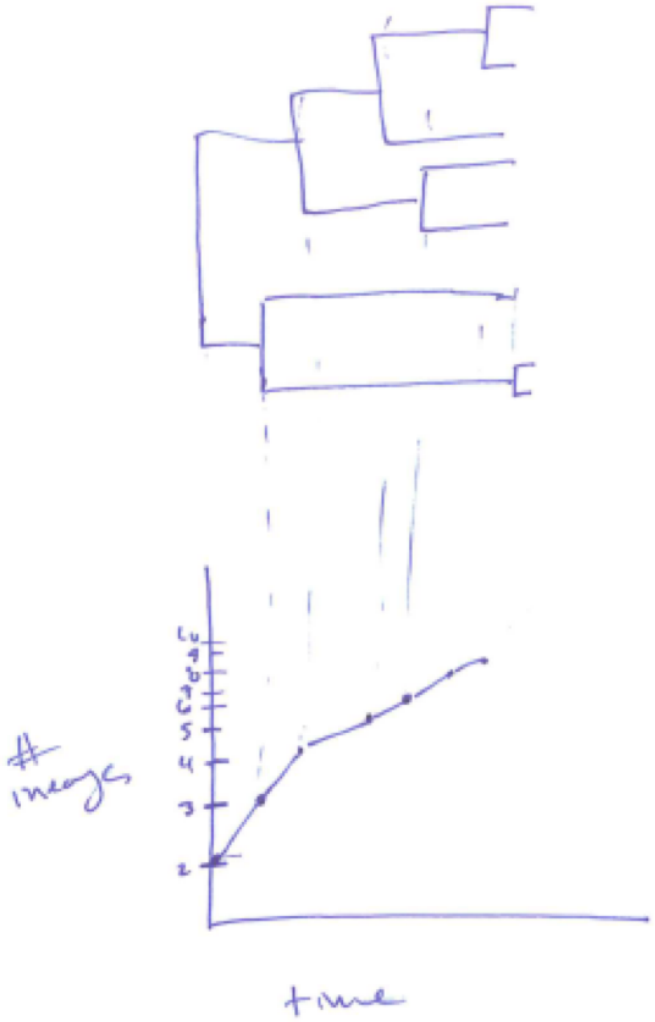

Figure 8: Figure 10.8. Lineage-through-time plot.

straight line on average (Figure 10.9A). By contrast, extinction should leave a clear signal in LTT plots because the probability of a lineage going extinction depends on how long it has been around; old lineages are much more likely to have been hit by extinction than relatively young lineages. We see this reflected in LTT plots as the "pull of the present" – an upturn in the slope of the LTT plot near the present day (Figure 10.9B). Incomplete sampling – that is, not sampling all of the living species in a clade – can also have a huge impact on the shape of LTT plots (Figure 10.9C). We will discuss LTT plots further in chapter 11, where we will use them to make inferences about patterns of lineage diversification through time.

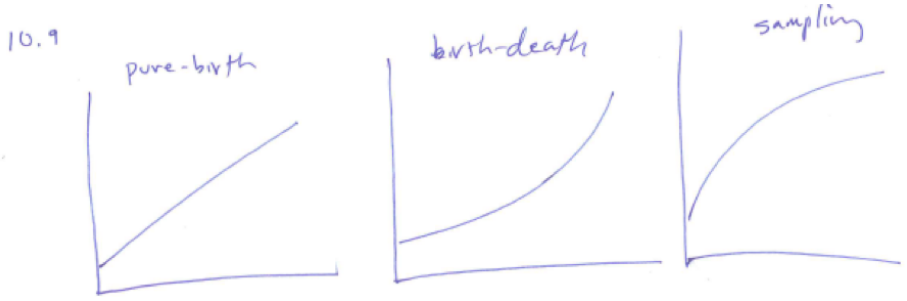

Figure 9: Figure 10.9. Example LTT plots.

Summary

In this chapter, I introduced birth-death models and summarized their basic mathematical properties. Birth-death models predict patterns of species diversity over time intervals, and can also be used to model the growth of phylogenetic trees. We can visualize these patterns by measuring tree balance and creating lineage-through-time (LTT) plots.