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

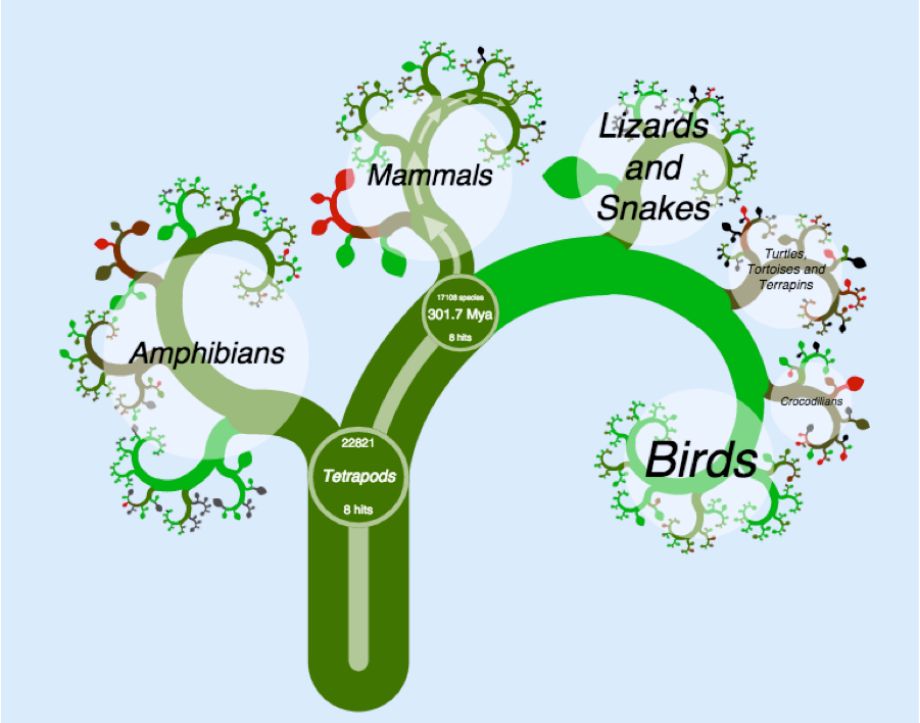

Figure 1: Figure 1.1. A small section of the tree of life showing the relationships among tetrapods, from onezoom (Rosindell and Harmon 2012). Arrows lead to you.

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

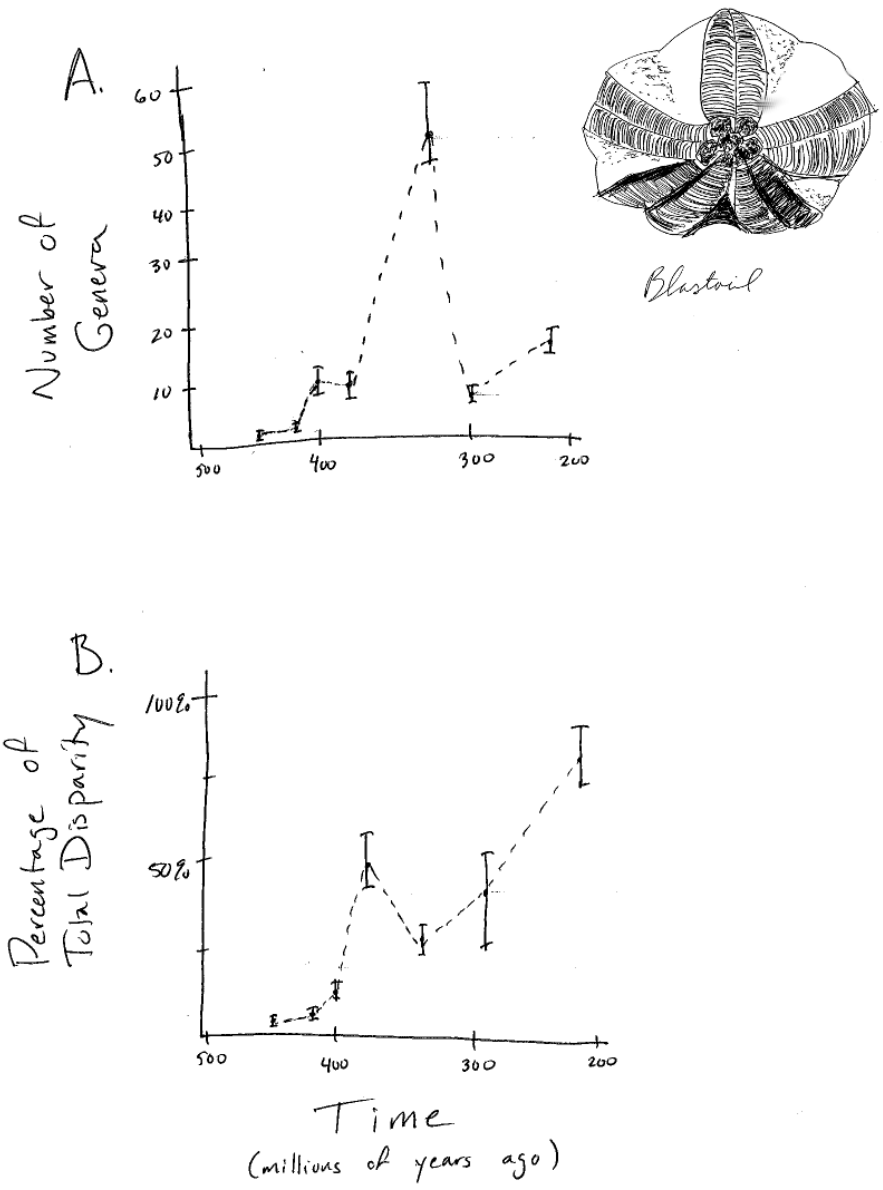

Figure 2: Figure 1.2. Diversity and disparity in the fossil record for the Blastoids. Plots show A. diversity (number of genera) and B. disparity (trait variance) through time. Taken from (Foote 1997).

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 pH = 0.5. The binomial test uses the number of "successes" (we will use the number of heads, 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 pH = 0.5; our alternative, then, is that pH takes some other value: pH 0.5.

To carry out the test, we consider the distribution of our test statistic (the number of heads) under our null hypothesis (pH = 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 pH = 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 pH $>$ 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-

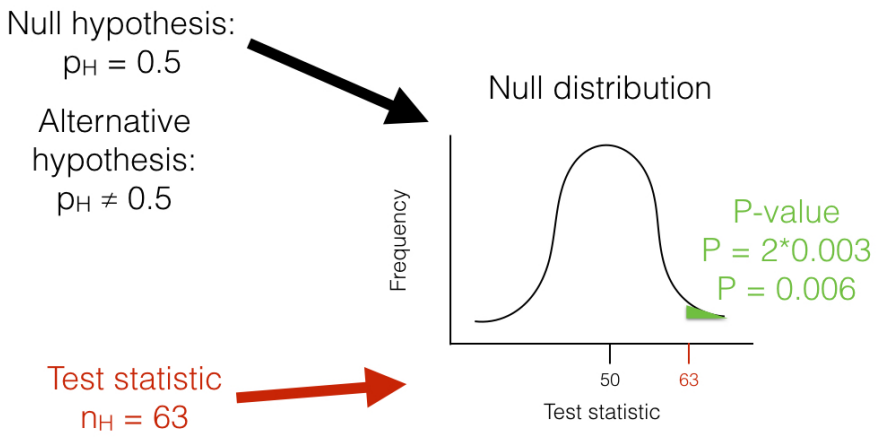

Figure 1: Figure 2.1. The unfair lizard. We use the null hypothesis to generate a null distribution for our test statistic, which in this case is a binomial distribution centered around 50. We then look at our test statistic and calculate the probability of obtaining a result at least as extreme as this 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, , which defines the accepted rate of type I errors. For example, if $= 0.05$, 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)

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

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

(eq. 2.3)

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

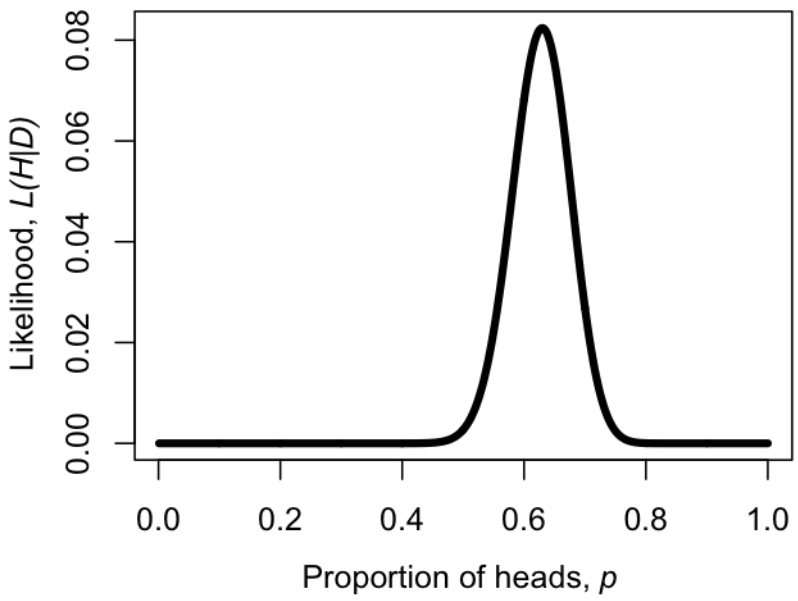

Figure 2: Figure 2.2. Likelihood surface for the parameter p, given a coin that has been flipped as heads 63 times out of 100.

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

$$\hat{p}$$

, is at p = 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 through differentiation. This problem is much easier if we work with the log-likelihood rather than the likelihood itself (note that whatever value of p that maximizes the likelihood will also maximize the log-likelihood, because the log function is strictly increasing). So:

(eq. 2.4)

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

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)

$$\frac{d \ln L}{dp} = \frac{k}{p} - \frac{(n-k)}{(1-p)}$$

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

(eq. 2.6)

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

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 $\hat{p} = 0.63$. 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 $p = 0.5$? 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, compared to the former (simple) model; the simple model is a special case of the complex model when p = ½.

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 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 chi-squared 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 chi-squared 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 = 0.5 as:

(eq. 2.8)

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

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

(eq. 2.9)

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

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, $P(H) = 1$ and $P(D|H) = P(D)$, so that $P(H|D) = 1$. 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 must be estimated. We can set a uniform prior between 0 and 1 for p, so that $f(p) = 1$ for all p in the interval [0,1]. We can also write this as "our prior for p is U(0,1)". Then:

(eq. 2.21)

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

Next we note that $P(D|H)$ 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)

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

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

(eq. 2.23)

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

We can compare this posterior distribution of our parameter estimate, p, 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

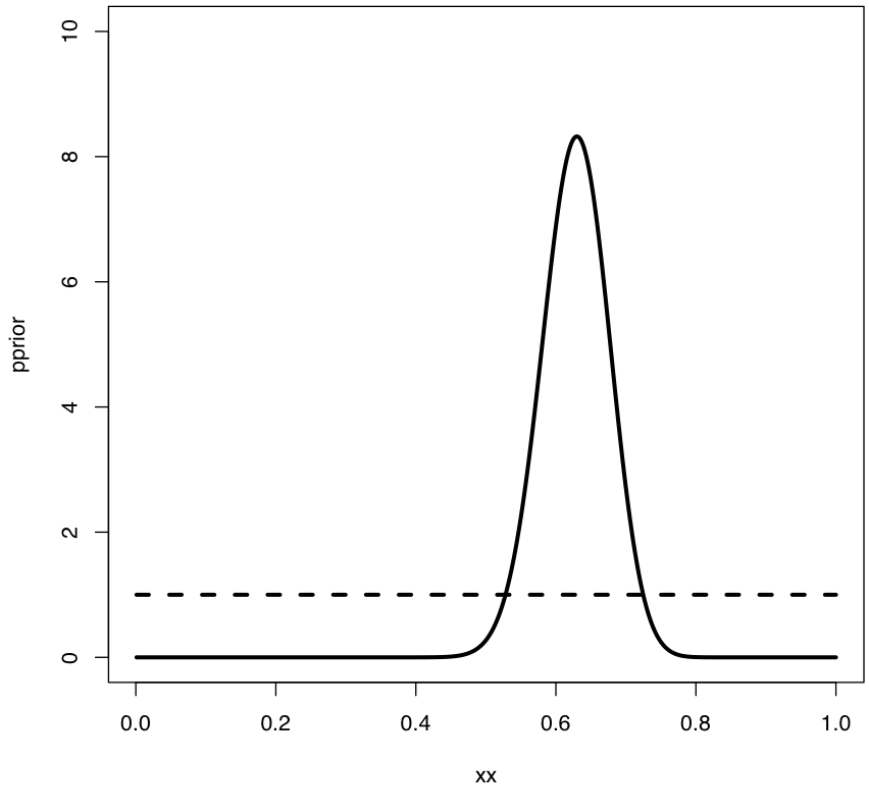

Figure 3: Figure 2.3. Bayesian prior (dotted line) and posterior (solid line) distributions for lizard flipping.

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)

$$a_1 = \frac{P(p')}{P(p)}$$

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)

$$a_2 = \frac{Q(p'|p)}{Q(p|p')}$$

c. The likelihood ratio.

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

(eq. 2.26)

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

4. Multiply.

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

(eq. 2.27)

$$a = a_1 \cdot a_2 \cdot a_3$$

5. Accept or reject.

Draw a random number x from a uniform distribution between 0 and 1. If $x < a$, 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)

$$a_1 = \frac{P(p')}{P(p)} = \frac{1}{1} = 1$$

Similarly, because our proposal distribution is symmetrical, $Q(p'|p) = Q(p|p')$ and $a_2 = 1$. That means that we only need to calculate the likelihood ratio 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)

$$a_3 = \frac{P(D|p')}{P(D|p)} = \frac{0.064}{0.068} = 0.94$$

We can now calculate $a = a_1 \cdot a_2 \cdot a_3 = 1 \cdot 1 \cdot 0.94 = 0.94$. 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 a, 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

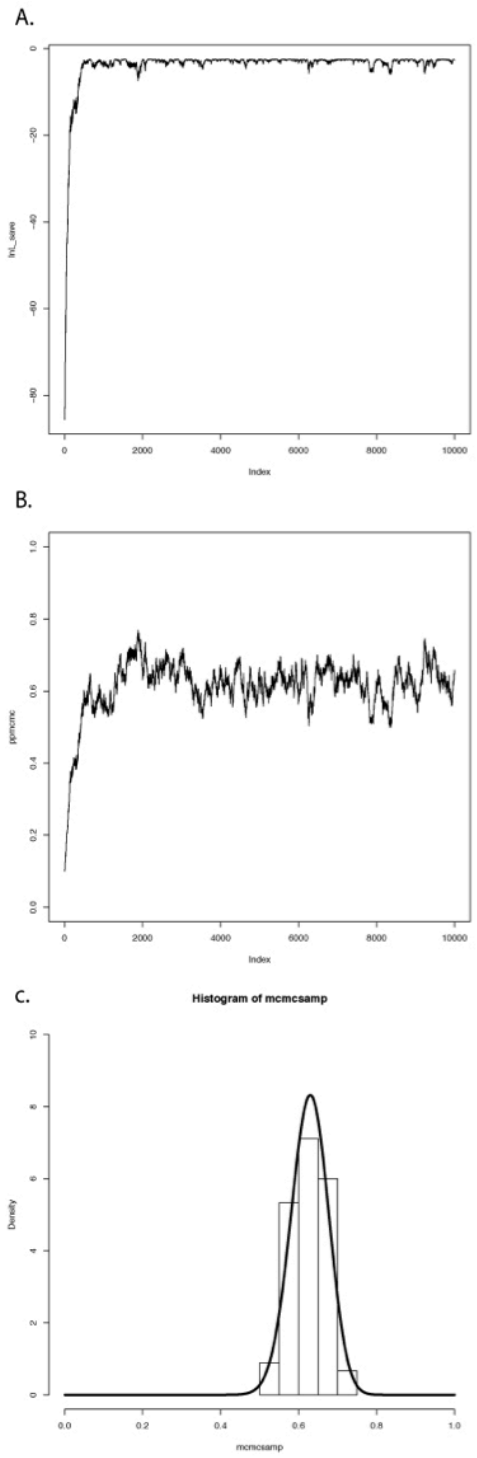

Figure 4: Figure 2.4. Bayesian MCMC from lizard flipping example.

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 $p = 0.63$. 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 $p = 0.5$, and model 2, where $p = 0.63$. 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.

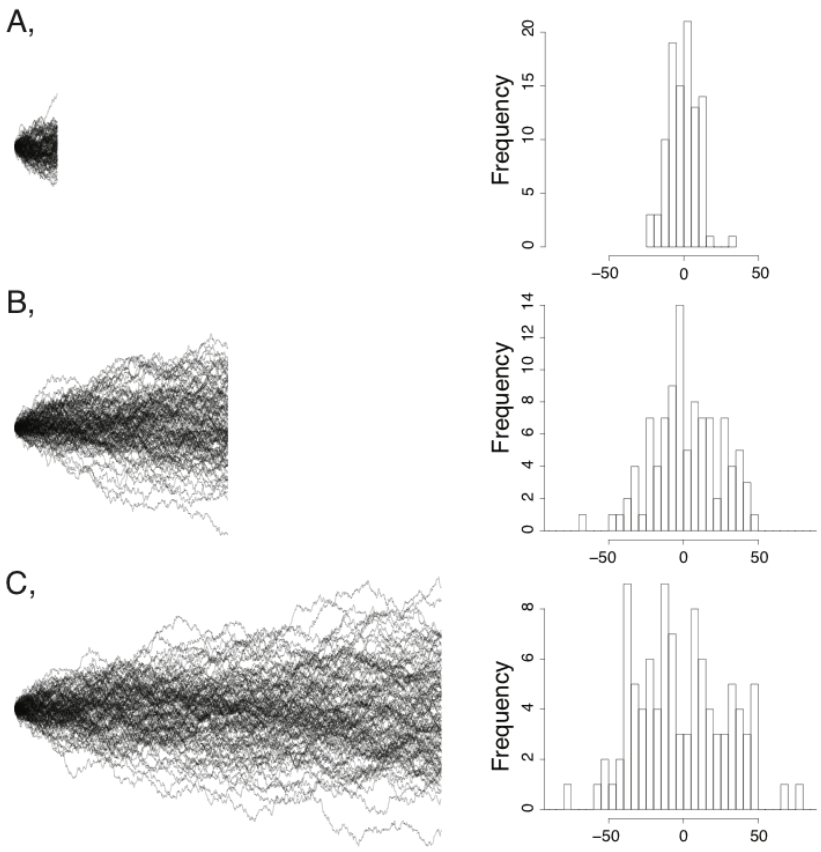

Figure 1:

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 $\sigma^2 t = 100$.

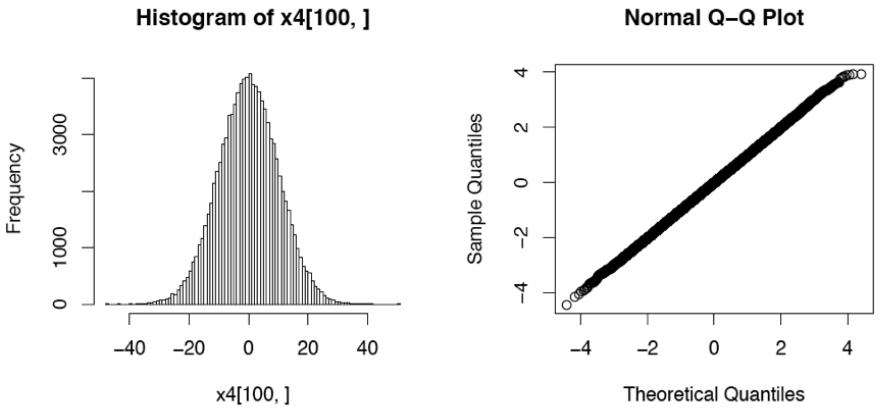

Figure 2:

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.

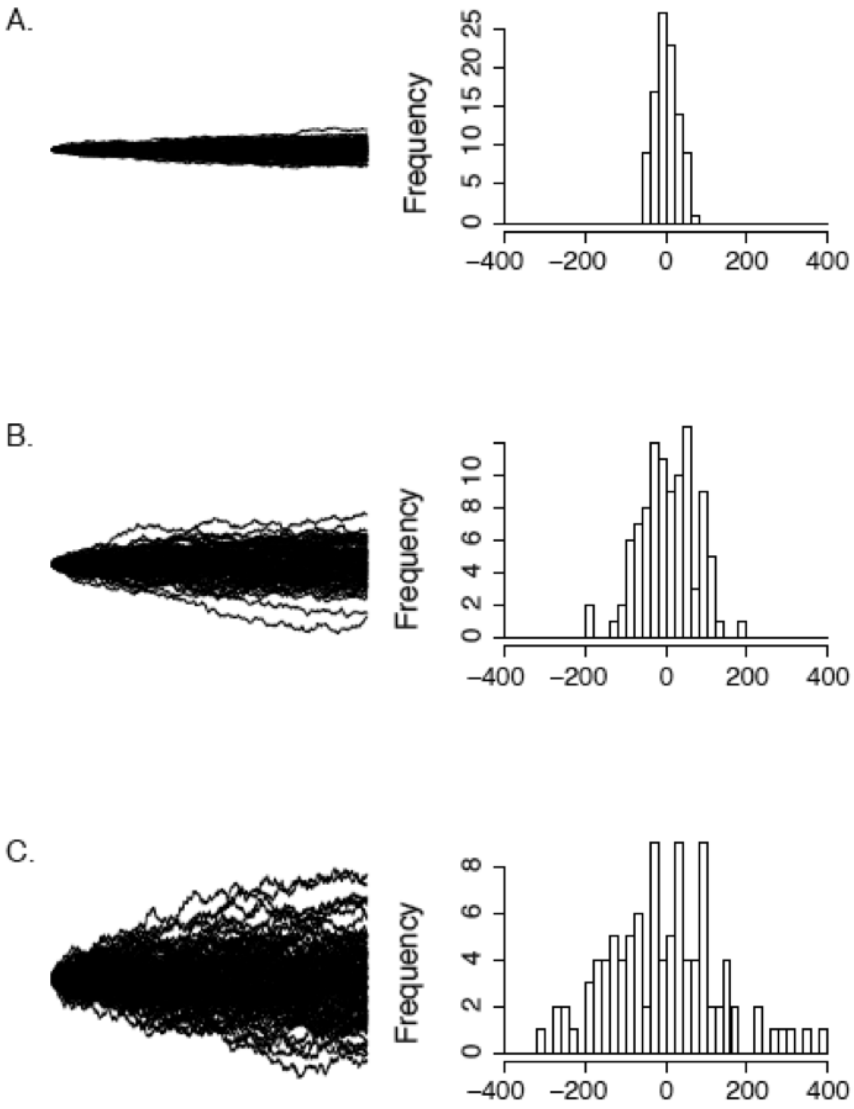
A.
Frequency
0 5 10 15 20 25
−400 −200 0 200 400
B.
Frequency
0 2 4 6 8 10
−400 −200 0 200 400
C.
Frequency
0 2 4 6 8
−400 −200 0 200 400

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)

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

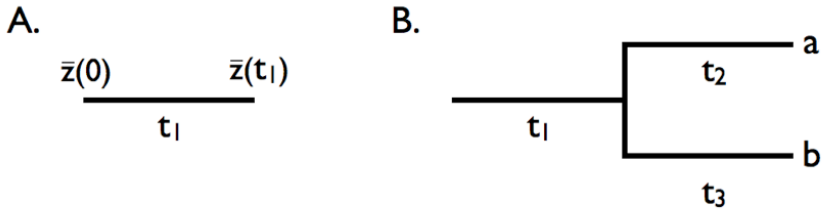

Figure 4:

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.

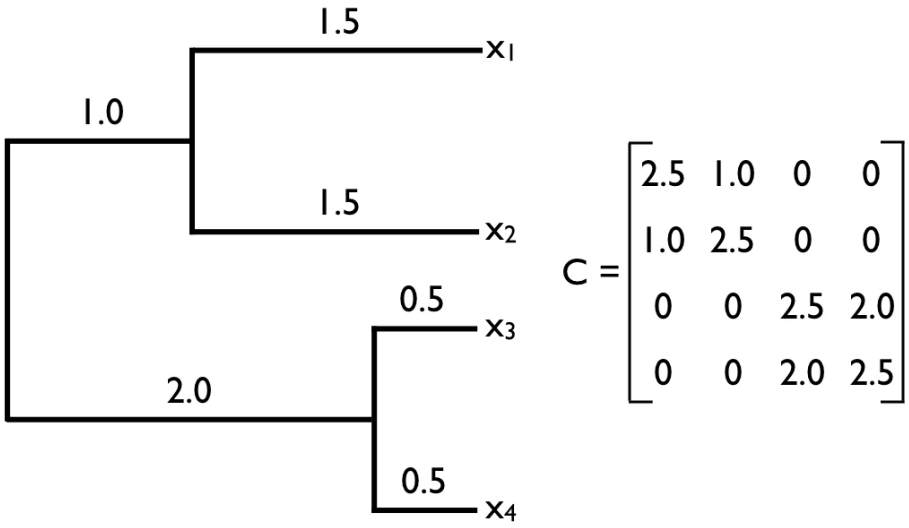

Figure 5:

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.