László Zsolt Garamszegi

Editor

Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology

Concepts and Practice

Springer

Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology

László Zsolt Garamszegi
Editor

Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology

Concepts and Practice

Springer

Editor
László Zsolt Garamszegi
Department of Evolutionary Ecology
Estación Biológica de Doñana-CSIC
Sevilla
Spain

Cover figure: Miklós Laczi

Printed on acid-free paper

An Online Practical Material (OPM)
for this book is available at:
http://www.mpcm-evolution.com
This website replaces the previous
website: http://www.mpcm-evolution.org.

Foreword

Humans and chimpanzees share around 99.8 % of their evolutionary histories, and nearly that much identity in the sequences of their genes, and yet most people are happy to lock the one up behind bars, throw it bits of food and watch it have sex, while protecting the other's rights with an exhausting array of laws, and granting it unique access to God's goodwill.

The surprise and even disorientation this image elicits goes to the heart of why comparative biology is fundamental to studying evolution and adaptation. Any two species that share that much of their histories are bound to be similar in many respects, as indeed humans and chimpanzees are. But we can also see in these two lineages the magisterial workings of natural selection, which has sculpted them in strikingly different ways since they parted company somewhere in Africa, perhaps 6–10 million years ago.

And that is the task that comparative methods must confront: how to separate what is shared from what is newly evolved, because it is the twists and turns of the latter that reveal how an evolving lineage has responded—and in most cases adapted—to its circumstances when compared to another evolving lineage whose circumstances differ. Remarkably, this simple truth has only become widely appreciated among evolutionary biologists in the last 25 years or so, and even today some recalcitrant troglodytes refuse to acknowledge it, or perhaps more charitably, the word has yet not filtered as far as their caves, or in one or two cases I know of, desert hideouts.

But happily, these people are an ever-shrinking minority whose eventual extinction no one will record on a Red List, and comparative methods are now widely used, having jumped from their historical origins in studying morphological and behavioural evolution, to virology, biogeography, ecology, gene and genome evolution, acoustics and perception, and to ever-finer analyses of the genetic basis of variation in organismic traits. Anthropologists use it, as do sociologists, and comparative methods have even been used to track down a deadly dentist who infected his patients with HIV.

But touting the hegemony of comparative methods is to adopt a narrow view. Their real power lies in combining information on the phylogenetic relationships among a group of species with data on the traits and behaviours of those species. Then, in combination with some hunches about the way evolution proceeds, hunches that take the form of statistical models of evolution, comparative methods

grant their users the power to reconstruct what the past was like, and then to test competing ideas about how that past gave rise to the present. When used properly, and with imagination, comparative methods make it possible to re-play the tape of evolution under different scenarios, giving credence to some scenarios over the others based on how well they explain the present.

It is to giving researchers the tools to do just this that László Zsolt Garamszegi's edited volume is devoted. It is a timely work as the past 10 years or so have witnessed useful ferment in the field as ever more fine-grained, large and detailed datasets and phylogenies become available, and statistically minded researchers have responded with ever more flexible methods. Garamszegi has organized the chapters so that readers are first prepared for and then gently eased into these various approaches. *Modern Phylogenetic Comparative Methods* will also have an Internet presence, allowing Garamszegi to keep users apprised of new developments and interpretations.

These are both useful strategies because different methods carry implicit controversies concerning the ways that evolution is assumed to proceed and what its patterns mean, and sometimes divide their followers. For instance, Brownian motion is still the dominant null-model of most investigations of continuously evolving traits, but where one camp adopts a catholic view, treating departures from a Brownian background as a statistical problem with multiple possible evolutionary causes, including varying rates of evolution and episodic bursts of change, others adopt a more Calvinist tone, insisting on 'stabilising selection' or attraction to a mysterious niche-optimum, as the cause—and their models admit no other view.

But these opposing views are—or should be—minor irritants in what is otherwise a scientific field in rude health, and one that in many respects should become the subject of young Ph.D. students in the philosophy of science. The old-guard Feyerabend-esque naysayers who cling to the desperate belief that science is just the province of who can shout loudest, and most effectively corrupt and coerce others, all in pursuit of their favourite myths, should take stock of the field of comparative biology: combative, and yes, often petty and self-serving, it has in these past 25 years or produced a steady, even if sometimes stumbling, triumph of the scientific method applied to this particular outpost of the field of evolution.

Mark Pagel

February 2014

Preface

Many evolutionary biologists are concerned with the tremendous amount of diversity of species, and their phenotypes, that we can currently observe in nature, and one of the most challenging tasks is to find evolutionary explanations for such interspecific differences. The philosophy of comparing attributes of different species that have undergone different selection regimes has heavily dominated the way we think about evolution since the days of Darwin. Subsequently, a vast number of studies have interpreted inter-specific variations in species-specific traits in the light of parallel variations in certain environmental or social factors.

However, a significant paradigm shift in the application of comparative methods occurred in the 80s, mostly from the influential work of Felsenstein in 1985, when it was recognized that patterns of inter-specific variations cannot be interpreted without taking into account the underlying common descent of species that delineates certain evolutionary constraints and also leads to similarity between species' phenotypes. Statistically, the effect of phylogeny can be regarded as a confounding factor that violates assumptions about non-independence of the unit of analysis, and that potentially introduces spurious correlations across traits. The development of the independent contrast method caused a flourish in the literature, in which most comparative papers adopted the method as an efficient way to get rid of these unwanted effects of phylogeny.

Modern phylogenetic methods go well beyond of this simple task of achieving a simple statistical control for phylogeny. They rather treat the evolutionary history of species as an interesting phenomenon on its own, that allows tracing character states back through time along series of ancestral states. The real value of modern phylogenetic methods is, therefore, the capacity to examine biological diversity in the light of the phylogeny, a perspective that opens up horizons for making inferences about where, when, and how traits have changed over an evolutionary time scale. The phylogenetic comparative framework by today has grown to address a large number of fascinating questions about the correlated evolution of traits, phylogenetic signals in interspecific data, ancestral states, the mode of evolution, evolutionary rates, alternative evolutionary mechanisms, speciation and diversification and between-species interactions. Most of the methods are available for both discrete (like mating system) and continuous (like body size) characters, and have already started to spread into disciples other than evolutionary biology (e.g., anthropology, genomics, linguistics, law, and sociology). Furthermore, given that

the phylogenetic comparative approach essentially offers a general framework for studying hierarchically structured data of any kind, one must assume that the exploitation of the underlying toolbox it provides has lagged behind its inherent potential.

In the current state of evolutionary biology, when phylogenies and interspecific data are accumulating at an enormous speed, it is becoming crucial that practitioners are armed with a diversity of comparative methods that help them make inferences from such data. However, overseeing these statistical tools becomes particularly challenging because of the richness and mathematical complexity of the relevant literature. In that situation, a secondary source, in which the primarily literature is brought into the attention of the user community in a consumable way may enhance the statistical integration of the discipline. This book has been assembled under this motivation, and aims at providing descriptions about the most recent phylogenetic methods for evolutionary biologist.

There are several other, very useful books on similar topics. Our book is not only novel because it revolves around the most recent developments, but also because this is an edited volume comprising contributions from several authors, each being expert on his/her respective subfield. This extensive collaborative project may, hence, offer a broad focus on a diverse array of topics and perspectives, which could not be covered efficiently by a handful of specialized authors only. The contributors to this book have been working in different fields of evolutionary biology for many years, under the ultimate aim of solving important biological questions. While fulfilling this enthralling mission we are often confronted with the task of broadening existing approaches and exploring new advances in the comparative study of diverse taxonomic systems and communities. In the light of this scientific background and our experiences obtained during teaching and in various statistical courses, we felt that a textbook was needed to provide a broad overview in the field of phylogeny-based evolutionary biology to our students and less experienced colleagues. Therefore, we wanted to compile a wide range of different perspectives and practices in the phylogenetic comparative method: from an introduction to the topic, through the diversity of statistical designs that can powerfully incorporate phylogenetic information, to more enhanced applications that offer studying evolutionary mechanisms. We must note, however, that due to various constraints, we were not able to review the entire literature that might be relevant. We hope that we will be able to adjust for such a shortcoming in the future.

Another extra value we offer is the accompanying online resource (available at http://www.mpcm-evolution.com), where we wish to post and permanently update practical materials to help embed methods into practice. As Online Practical Material (OPM), we will provide tutorials, example files and the underling statistical scripts, with which the users will be able to apply the presented methods to their own data and scientific questions. New approaches appear like mushrooms in the forest, with some of them being implemented in practice at a rapid pace. It is also becoming very common that more than one statistical approach is available for the same evolutionary problem. Being able to select from alternative approaches also allows researchers to gain better comprehension of biological diversity prevent

in natural systems. Therefore, it is vital to keep these methodologies updated and accessible for the broad user community, and the attributes of an online surface can be fruitfully exploited in that direction.

Statistics is not without mathematics, and the derivation of certain formulas is unavoidable for the appropriate argumentations. In this book, we constrain ourselves to present only the key mathematical formulas that are necessary for understanding the philosophy of different approaches. Doing so should make the reasoning accessible to a broad readership. For those who are interested in the mathematical details, we provide pointers to the primary literature. And for those not interested in the mathematical details, do not fear. The equations herein should be treated as resources if you decide to incorporate these methods into your research.

The quality of this book in large part resides on inputs from several expert reviewers, who provided valuable comments on earlier versions of chapter manuscripts during their review phase. Along that line, we are extremely grateful to David Bapst, Roger Benson, Kierstin K. Catlett, Natalie Cooper, Thomas Currie, Pierre de Villemereuil, Joe Felsenstein, Sive Finlay, Rob Freckleton, Jesualdo Fuentes-Gonzalez, Alejandro Gonzalez-Voyer, Tatiana Giraud, Alan Grafen, Randi Griffin, Jarrod Hadfield, Thomas Hansen, Lam Si Tung Ho, Elisabeth Housworth, Antony Ives, Jason Kamilar, Tom Kraft, Oriol Lapiedra, Jessica Light, Graeme Lloyd, Rafael Maia, Emilia Martins, Nick Matzke, Mark A. McPeek, Magdalena N. Muchlinski, Shinichi Nakagawa, Chris Nasrallah, Charles L. Nunn, Christopher E. Oufiero, Emmanuel Paradis, Sandrine Pavoine, Matt Pennell, Pedro Peres-Neto, Samantha Price, Liam Revell, Aaron Sandel, Holger Schielzeth, William Shipley, Graham Slater, Jeet Sukumaran, Matthew Symonds, Gavin Thomas, Chris Venditti, Jamie Caroline Winternitz and Derrick Zwickl for their constructive assistance in the evaluation process.

Although the list of authors appears male biased, this is completely unintentional. Female scientists were also invited to write chapters, but regretfully they could not make contribution due to various and completely understandable reasons.

Seville, Spain   László Zsolt Garamszegi

Contents

Part I
Introduction

Chapter 1
An Introduction to the Phylogenetic Comparative Method

Emmanuel Paradis

Abstract The phylogenetic comparative method (PCM) has an important place in evolutionary biology. This chapter aims at giving an overview on some selected topics. We first review briefly some important historical milestones including some early contributions and the relationships of comparative methods with phylogenetics. Some fundamental points on statistical inference, adaptation, and causality are then discussed. We also discuss briefly the application of the PCM to anthropology and conclude with some perspectives on its future development and applications.

1.1 Introduction

> A comparison of apples and oranges occurs when two items or groups of items are compared that cannot be practically compared. … However, apples are actually more closely related to pears (both are rosaceae) than to oranges.
>
> —Wikipedia[1]

The phylogenetic comparative method has undoubtedly been one of the most important phenomena of evolutionary biology during the last few decades. Comparative methods exist in many fields such as anthropology (Bock 1966), law (Kiekbaev 2003), linguistics (Forster et al. 1998), and evolutionary biology (Harvey and Pagel 1991). The concepts and uses of these different comparative methods vary widely. Since the present book is specifically concerned with biological evolution, it is thus useful to define our subject.

---

[1] http://en.wikipedia.org/wiki/Apples_and_oranges

E. Paradis (✉)
Institut de Recherche pour le Développement, Montpellier, France
e-mail: emmanuel.paradis@ird.fr

We may define the *comparative method* as an analytical approach based on the comparison of different objects with the aim to elucidate the mechanisms at the origin of their diversity. From this, we can define the *phylogenetic comparative method* as the analytical study of species, populations, and individuals in a historical framework with the aim to elucidate the mechanisms at the origin of the diversity of life.

It is important to note that the phylogenetic comparative method (PCM) is distinct from but not independent of *phylogenetics*, the study and reconstruction of the historical relationships among species. For instance, in linguistics or in anthropology, the goal of the comparative method is the historical reconstruction of spoken languages or of human cultures (e.g., Forster et al. 1998, with some historical references therein; see Sect. 1.7 below).

The goal of this chapter is to give a general introduction to the PCM by examining some topics. The next section presents the main historical milestones of phylogenetics and the comparative method—since both have been tightly linked through their history. The following sections give some essential elements on statistical inference of evolutionary processes with comparative data. The last two sections aim to put the PCM in a broader perspective by looking at its relationships with anthropology and speculating about some of its current advances and its future.

1.2 History of Phylogenetics and the Comparative Method

1.2.1 Early Developments

In the nineteenth century, trees were essential graphical tools for the development of evolutionary ideas. Lamarck (1809) used a downward-growing tree to represent the relationships among the main groups of animals with a caption indicating that this "table displays the origin of the different animals." History was central in Lamarck's argumentation: "A strong reason prevents us to identify the changes that have successively diversified the animals as we know them today: we have never witnessed these changes." This could be taken as a manifesto of today's comparative method in evolutionary biology.

Cuvier was Lamarck's great rival and strong opponent to the idea of evolution. However, Cuvier acknowledged that species are more or less closely related so that they can be classified in a hierarchical system and that different characters of these species are relevant at different levels, especially through his gradual characters (*caractères gradués*, Cuvier 1798). In spite of his backward ideas on fixism (Laurent 1986), Cuvier had a profound impact on comparative anatomy through the numerous illustrations and drawings included in his books—which could appear in a modern textbook on evolution after updating the captions.

It has been widely appreciated that Darwin (1859) used a phylogenetic tree as the only figure in the *Origin of Species*. He also used comparative data to support several of his points; for instance, "Genera which are polymorphic in one country seem to be, with some few exceptions, polymorphic in other countries, and likewise, judging from Brachiopod shells, at former periods of time." Thus, Darwin characterized the patterns of diversity in space and in time and also used these facts to infer the processes of diversification of species: "…the larger genera also tend to break up into smaller genera. And thus, the forms of life throughout the universe become divided into groups subordinate to groups."

The end of the nineteenth century has witnessed the wide acceptance of the idea of evolution, particularly with the contributions of Haeckel, the father of phylogenetics: "For the purpose of constructing a hypothetical genealogical tree of the Radiolaria, as of all other organisms, three sources of information are open to us, viz., palæontology, comparative ontogeny, and comparative anatomy." (Haeckel 1887).

During the first half of the twentieth century, phylogenetics and biological comparative studies have followed separate paths. The discovery of the physical support of heredity (genes, chromosomes, and later DNA) led scientists to focus their interest on the genetic mechanisms of evolution. Fisher (1930) certainly best illustrates this change of paradigm where history was less important than previously thought: "For mutations to dominate the trend of evolution it is thus necessary to postulate mutation rates immensely greater than those which are known to occur…" For Fisher, mutations could not explain evolutionary novelties and we should rather look at other evolutionary forces such as selection or population structure to explain the diversity of life. At the same time, phylogenetics made fundamental contributions to evolutionary thinking. Paleontologists integrated phylogenetic ideas, mainly because of the historical nature of their data (Simpson 1944). Phylogenetic trees became the analytical tool of a school of systematists (cladistics), leading to the first numerical treatments of phylogenies (see a historical account in Felsenstein 2004).

1.2.2 Modern Developments

The late 1960s have witnessed some crucial turns. The development of statistical methods to reconstruct phylogenies from genetic data was a major step accomplished by Cavalli-Sforza and Edwards (1967). Because the approach they developed was statistical, it was possible to extend it to other kinds of data such as continuous characters. This next step was achieved by Felsenstein (1973) who proposed a method to calculate the likelihood of a tree for a set of continuous traits. The significance of this work was not obvious until the same author published a related method to calculate the phylogenetically independent contrasts (PICs), a major difference being that the calculations under this new method could be done with a hand calculator (Felsenstein 1985).

Until the 1970s, comparative biology developed its statistical tools independently of phylogenetic or historical ideas. Comparative data from n species used to be analyzed with standard statistical methods, assuming that they were n independent observations. This separation between comparative biology and the historical dimension of biological evolution seems surprising when considering that the idea of evolution, and particularly adaptation, was at the heart of most comparative studies (Clutton-Brock and Harvey 1979).

The 1980s can be seen as the golden age of PCMs when a great variety of methods were published (reviewed by Pagel and Harvey 1988). Two important papers were published in the same year: Cheverud et al. (1985), who proposed an approach based on auto-regression including the possibility to account for intraspecific variation, and Felsenstein (1985), already cited. A few years later, Grafen (1989), in a very rich and dense paper, proposed the use of generalized least squares (GLS) to derive a method now widely known as the phylogenetic generalized least squares (PGLS). Gittleman and Kot (1990) further developed the use of auto-correlation functions to assess phylogenetic signal in diverse settings, including using taxonomic levels when a phylogeny is not available. Nevertheless, the power of these methods to infer evolutionary models and parameters was not yet fully acknowledged, and the view that phylogeny was a confounding effect in comparative analyses still prevailed: "Confounding effects of phylogeny and other variables may lurk behind any comparative relationship, and they must be removed or controlled prior to considering adaptive arguments." (Pagel and Harvey 1988).

During the 1990s, the developments of the previous decade were confirmed and strengthened. Lynch (1991) made a link between models of quantitative genetics and phylogenetics. He developed a method to partition the variance of a trait into an environmental and a phylogenetic component. Importantly, the same decomposition can be done for the covariance between two traits, thus providing a formal way to quantify the historical component of the link between two characters. New methods were proposed for the analysis of discrete traits (Pagel 1994; Grafen and Ridley 1997). In two important papers, Hansen and Martins (1996) and Martins and Hansen (1997) showed how GLS can be used to address evolutionary questions beyond the basic Brownian motion model.

In the 2000s, some efforts were given to issues left temporarily aside such as fitting more complicated models combining continuous and discrete traits (Paradis and Claude 2002; Felsenstein 2005; Hadfield and Nakagawa 2010) or combining interspecific and intraspecific data (Felsenstein 2008; Garamszegi and Møller 2010; Stone et al. 2011; see Chap. 7). At the end of the decade, three papers by Revell (2009), Jombart et al. (2010), and Pavoine et al. (2010) defined a general framework for multivariate statistical analyses in a phylogenetic context. The concept of phylogenetic signal has also attracted significant interest with the aim of clarifying previous ideas on phylogenetic confounding effect (Blomberg et al. 2003; Ollier et al. 2006; Pavoine et al. 2008; Münkemuller et al. 2012). At the same time, PCMs have achieved maturity with some generalizations such as the development of a Brownian model with variable parameters (O'Meara et al. 2006)
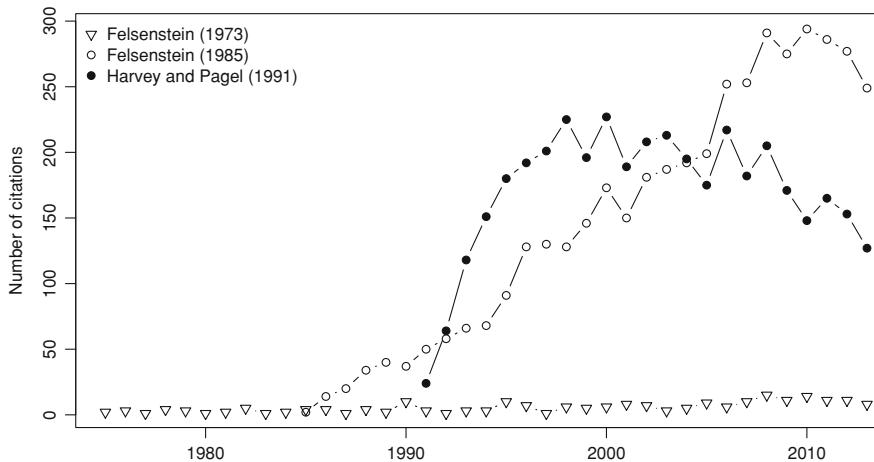
Fig. 1.1 The annual number of citations of three major contributions to the phylogenetic comparative method (*Source* Web of Science)

or the use of sophisticated model-fitting techniques such as Monte Carlo Markov chains (MCMC) to fit complicated models (Pagel et al. 2004; Pagel and Meade 2006; Hadfield and Nakagawa 2010).

Remarkable progress in phylogenetics also contributed to PCMs, particularly with the publication of more and more phylogenies, some of them being complete over a very large number of species (Bininda-Emonds et al. 2007; Smith et al. 2011; Jetz et al. 2012; see Chap. 3). Some methods have been developed to combine different sources of phylogenetic information in order to build trees for comparative analyses (Kuhn et al. 2011; Eastman et al. 2013; Thomas et al. 2013; see Chap. 2).

Figure 1.1 gives the number of citations of two earlier papers over the years together with another major contribution to the development of PCMs. After almost three decades, the range of applications of PCMs has grown to reach all branches of biological science: 6,533 citations of Felsenstein (1985) or Harvey and Pagel (1991) are found in 771 periodical titles. The PCM, through the development of a wide range of analytical tools, has contributed insights into many questions on evolution and the diversity of life.

1.3 The Covariance Structure of Comparative Data

A central issue with comparative data is the non-independence of observations. A similar problem is found in other fields such as geography (Cliff and Ord 1981), climatology (Tiao et al. 1990), ecology (Legendre 1993), or medical research (Houwing-Duistermaat et al. 1998).

Variables (Traits)
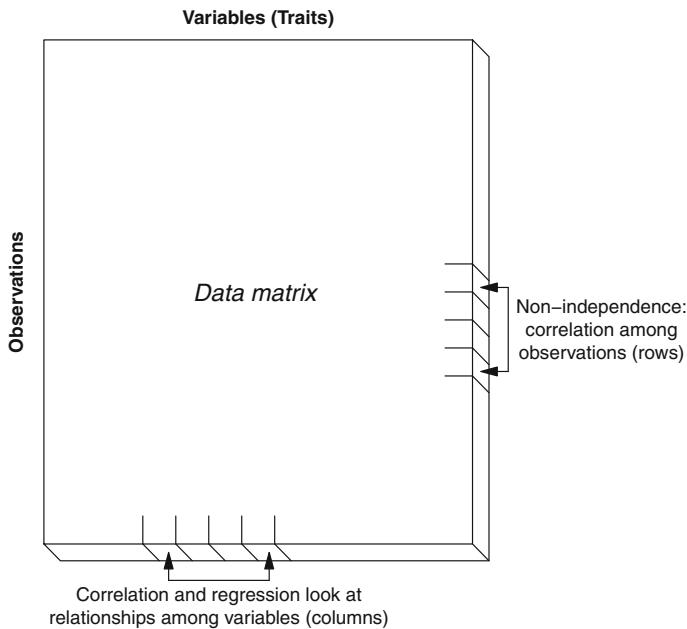
Fig. 1.2 A general depiction of a data set

In a very general way, a data set can be arranged in a matrix where the rows are the observations (individuals, populations, species, sequences, cells, etc.) and the columns are variables (size, area, nucleotide sites, RNA transcription levels, etc.). Data analyses seek for relationships among the columns of this matrix, and common statistical methods assume that the rows are independent observations; in other words, the values observed on a given row are not affected by the values at others (Fig. 1.2).

To statistically handle non-independence of observations, a general approach is to assume that two observations (rows), say i and j, are related through a covariance parameter denoted as σ_{ij}^2. This parameter specifies the strength of the relation between the values of the same variable (column) observed for these two observations. The way this parameter enters in the analyses depends on the method used, the kinds of variables, and the question asked. The covariance parameters are usually arranged in a symmetric matrix with the diagonal elements equal to the variances and the off-diagonal elements being the covariances (see Chap. 5). This matrix has n rows and n columns and so contains $n(n-1)/2$ off-diagonal elements.

There are many ways to define the values of σ_{ij}^2: they may be all equal or not, they may follow a specific distribution or may be related to another variable, they may be fixed or estimated from the data, etc. For instance, with spatial data, it is common to use a covariance function related to geographical distance.

In the case of comparative data from several species, it is possible to calculate a priori the covariances among species traits if we know how these traits have evolved. For instance, if we assume that a trait has evolved under a Brownian motion (BM) model, these covariances can be calculated from the phylogenetic tree linking these species without observing the trait itself. The PIC and PGLS methods were directly derived from this assumption. Both methods are identical though they are computationally very different (Blomberg et al. 2012). PGLSs directly use the covariances by calculating a correlation matrix among observations, which is simply a covariance matrix scaled to have values between -1 and 1 (see Chap. 5).

1.4 Statistical Inference

The covariance matrix is at the core of most PCMs. Fortunately, it is possible to calculate it for other models of trait evolution, in particular for the Ornstein–Uhlenbeck (OU) model which is appropriate to model evolution of traits under constraints (see Chaps. 13, 14 and 15). Using this and other models, it is thus possible to relax the assumption underlying the BM model. This feature of PCMs allows us to go beyond the paradigm that phylogeny is a confounding effect that must simply be corrected (see Rohlf 2006).

We can illustrate this point with a small simulation exercise. Taking the phylogeny in Fig. 1.3, we simulate two independent traits that evolve according to either a BM or an OU model. In this second model, the traits are constrained to evolve toward an optimal value with a strength controlled by the parameter denoted as α. The simulated traits were analyzed with two methods: a standard regression (assuming the species are independent) and a regression using the PICs calculated with the original phylogeny (which was thus assumed to be perfectly known). Table 1.1 shows the estimated rejection rates for both methods. Since both traits are independent, we expect these rates to be close to 5 %. The PIC-based analyses gave the correct answer with the data simulated from a BM model or from an OU model with a small value of α. On the other hand, for the large values of α, the PIC-based analyses had a high type I error rate, whereas the standard regression had a rejection rate close to 5 %.

Figure 1.4 shows the correlation matrices among the 20 leaves of the tree under different models of trait evolution. These matrices would be used in PGLS analyses (see Chap. 5). This shows clearly that an OU model with small α is close to a Brownian motion one, whereas when α is large, the observations are expected to be almost independent.

The critical point in a PCM-based analysis is to use the correct correlation structure among observations. The more distant the assumed correlation structure from the real one, the more biased the analysis will be. This property explains the statement that "in a comparative analysis a wrong phylogeny is better than no phylogeny at all" (Losos 1994; Martins 1996). Indeed, if the traits evolved on a
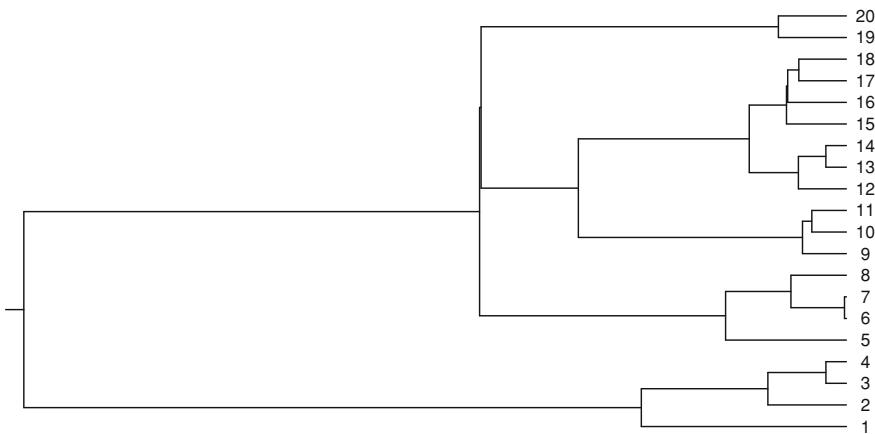
Fig. 1.3 A simulated phylogeny with $n = 20$

Table 1.1 Rejection rate of the test of correlation between two independent traits simulated on the phylogeny in Fig. 1.3 using the model indicated in the table

Simulated model	α	Standard regression	PIC regression
Brownian motion		0.396	0.051
Ornstein–Uhlenbeck	1	0.223	0.056
	10	0.065	0.120
	100	0.051	0.343

Simulations were replicated 10,000 times
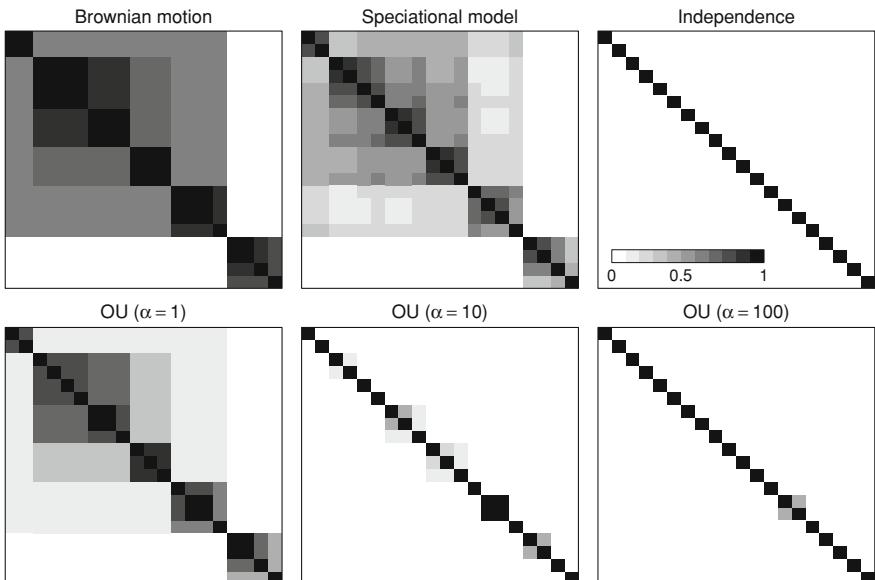
Fig. 1.4 Correlation matrices among the 20 tips of the tree in Fig. 1.3. The speciational model is one where change occurs only after a speciation event

phylogeny and another phylogeny is used for data analysis, the latter will result in a correlation structure closer to the correct one than assuming no correlation at all (i.e., independence of observations).

From the point of view of data analysis, one problem often encountered in published studies is that the phylogenetic correlation structure of the data is usually not assessed. This certainly comes from the view traditionally defended by most authors that only the phylogenetically controlled analyses are relevant. We now know that this can lead to wrong inference. This seems relatively easy to fix with tests of phylogenetic signal and model selection with information criteria such as the AIC (see Chap. 5). Remarkably, when different correlation structures are compared with real data, the Brownian motion model is rejected against more complex models such as the OU one (Whitney et al. 2011; Lapiedra et al. 2013).

1.5 Inferring Adaptation

Perhaps because of its history, the comparative method is most often used to infer adaptation. However, the scope and power of the PCM to reveal adaptation have been criticized several times over the years (e.g., Leroi et al. 1994; Martins 2000; Grandcolas et al. 2011). Such criticism is not really surprising: It has been discussed since a long time ago that characterizing the adaptative nature of a trait is a complicated endeavor (Bock 1959). Even the characterization of adaptation in viruses, which are far simpler than the organisms studied by most evolutionists, appears to be an arduous task (Pepin et al. 2010). The use of traits such as "habitat use" or "environment" with PCMs has been questioned because the analysis of such variables in a phylogenetic framework is meaningless (Grandcolas et al. 2011). On the other hand, it is hard to not consider these variables in evolutionary models since the assessment of the adaptive value of a trait cannot be separated from extrinsic variables such as habitat, resources, or climate (Bock 2003; Losos 2011; Watt 2013).

Some recent developments in PCMs provide a solution to the limitations underlined by the critiques cited above. As we have seen in the previous sections, the PCM does not simply aim at correcting for phylogenetic dependence or inferring repeated evolution of the same trait in different lineages, but rather to provide tools to analyze comparative data in a historical framework, and this includes fitting complex models of trait evolution that can handle various complications of the study design. For instance, some methods make possible to analyze several traits that evolve under different models: Bartoszek et al. (2012) developed a multi-trait model where traits can evolve following different processes of BM or OU.

One situation illustrated by Losos (2011) is the one of "incomplete convergence." Convergence toward similar phenotypes among distantly related species is often viewed as evidence for adaptation. However, adaptive evolution can proceed in different ways in different groups, and the patterns thus produced are likely to be

masked or obscured by other variables (such as the taxonomical background). Losos (2011) gives the example of the head shape of lizards which is mainly related to phylogenetic relatedness; however, within distinct clades, some species evolved independently toward herbivory and share some similarities, but the convergence is incomplete as they retain their respective phylogenetic background. In this case, a standard comparative analysis will likely fail to characterize the limited convergence among species affected by similar selective forces. On the other hand, statistical methods in a historical framework, including models of trait evolution, are helpful to characterize such patterns of adaptation. The use of these and other recently developed models of trait evolution may help to solve this "paradox" of using non-heritable traits in comparative analyses.

1.6 Inferring Causality

Correlation is not causality, and PCMs do not escape this reality. In spite of the importance of causality in evolutionary theory (see Watt 2013, for a recent view), the application of PCMs is generally oblivious of this point. This has led to some debate about the applicability of the PCM in order to identify evolutionary mechanisms. The vast majority of publications do not elaborate much on correlation and causality in their predictions: A simple linear correlation is usually derived from the hypotheses under test.

   In general statistical inference, the causal relationship between two variables (say x and y) can be assessed if one of them can be controlled and then used as a predictor in data analyses. With comparative data, x and y cannot be controlled: They are evolving traits (or intrinsic variables) which are measured "on the species" (or they are extrinsic variables, like habitat, which cannot generally be controlled). Therefore, in the situation of a PCM with two variables, it is not possible to determine which regression (x on y or y on x) best describes the data. In other words, we cannot infer the causal relationship between these two variables.

   When three or more uncontrolled variables are analyzed, it is possible to assess alternative causal relationships among them with a method known as path analysis (Freedman 2009). This method considers explicitly the causal relationships among variables under alternative hypotheses. A causal relationship can be expressed as "the variation in y is caused by the variation in the value of x" and has the statistical consequence that the regression of y on x is meaningful. Under a given hypothesis of the causal relationships among variables, some regressions are meaningful, while others are not. Using a procedure called the d-sep, it is possible to test which hypothesis best describes the data (Shipley 2013). Santos and Cannatella (2011) and von Hardenberg and Gonzalez-Voyer (2013) proposed to extend the framework of path analysis to PCMs (see Chap. 8). In a traditional path analysis, the regressions are done assuming independence of observations. Therefore, it is straightforward to generalize this method to cases where the observations are not independent, using tools such as (P)GLS.

Causality in general, and in evolution in particular, is fundamental, but this is a difficult concept to apply in practice. Hopefully, future applications of PCMs will help to progress on this issue.

1.7 Phylogenetic Comparative Method and Anthropology

As mentioned above, the present book focuses on the uses of the comparative method in evolutionary biology. One reason for this restriction is that other scientists do not see the comparative approach in the same way than evolutionists do. For instance, Bock (1966) described the comparative method as follows (italics as original):

> It should be recalled at the outset that the primary objective of users of the comparative method is historical reconstruction. *What* history or *whose* history is by no means clear in the nineteenth century literature, and this question has hardly been resolved in recent controversy.

Thus, what anthropologists call "comparative method" seems close to what evolutionists call "phylogenetics." Mace and Pagel (1994) revisited this issue by introducing a phylogenetic approach to anthropology directly inspired from evolutionary biology. Considering the links between comparative biology and phylogenetics, their message does not differ radically from the one formulated 28 years before by Bock.

In practice, the application of PCMs in anthropology differs substantially compared to evolutionary biology. A remarkable difference is that with anthropological data, the historical sequence of changes in traits (cultures, political systems, etc.) is often recorded—at least more often than in biology. For instance, Lindenfors et al. (2011) studied changes in political systems in the world between 1800 and 2008. Using several variables, they built a score ranging between −10 (total autocracy) and +10 (full democracy) and measured transitions among these different scores. They showed that most political changes occurred from autocratic systems (with a peak around −6) toward democratic ones (with a peak around +8). However, one interesting point about this study is their comment with respect to the historical dimension of the problem:

> A reconstruction of democracy as a political system on a language phylogeny would almost certainly indicate democracy as the ancestral state for large sections of the phylogeny. However, since we have exact information of all transitions, we know this not to be true.

In biological terms, there is a trend (or directional evolution) from autocracy toward democracy so that the second system is the most widespread today among countries. If we ignore this historical trend, we would make wrong inference. Here also, we see that using the wrong model of evolution can be misleading. Similar situations can be found with evolutionary data; for instance, if a trend exists in the evolution of a trait (say, increase in body size), then ancestral inference will likely

be misleading if this trend is not taken into account (Grafen 1989). Anthropological data have other peculiarities, like the ubiquity of horizontal transfers (Borgerhoff Mulder et al. 2006), so that a full comparison between cultural and biological evolution would require a much longer discussion.

1.8 The Future of the Phylogenetic Comparative Method

The comparative method in biology has evolved through several centuries to reach its present status. Today, PCMs have attained a level of maturity and sophistication that the readers can appreciate in the chapters of this book. The directions of future progress are certainly multiple.

We have seen the importance of the evolutionary models in statistical inference with the PCM. Some researchers currently explore the possibility to analyze complex models with several variables and an explicit formulation of the relationships among them. Hadjipantelis et al. (2013) analyzed an evolutionary model of "function-valued traits" by combining dimensionality reduction and a "bagging" (bootstrap aggregating) procedure. Complex relational models fitted with structural equation models seem also a very promising approach for future comparative analyses (Chap. 8).

It has not been widely appreciated that some PCMs link a model of microevolution (random evolution through genetic drift or stabilizing selection) with the patterns of interspecific variation in a trait and thus with macroevolution. Further works in this direction will likely lead to some interesting investigations on evolutionary mechanisms.

During many decades, the fossil record has been considered as the only source of information about evolutionary change. This paradigm has been broken by two steps forward in evolutionary biology: phylogenetic methods which try to reconstruct the past from the present and observations of real evolutionary changes over recent years such as the spread of resistance alleles in pathogens. However, the divorce between paleobiology and PCMs does not seem natural, and several researchers try to reconcile them (Pennell and Harmon 2013). This task will surely be very difficult but considering the many contributions of fossils to evolutionary biology, this is worth the effort (Chap. 22).

A scientific discipline is sometimes judged by how it contributes to everyone's well-being. The PCM may well move very positively in this direction as comparative analyses could have concrete applications. Spreitzer et al. (2005) used simple comparisons between different plant and alga species to create new enzymes involved in photosynthesis by targeted mutations on "phylogenetic residues." The resulting enzyme is one with original characteristics, a kind of "phylogenetic chimaera." Such an approach of "phylogenetic engineering" may be promising to design new proteins or even new organisms based on predictions from evolutionary phylogenetic models.

Yan et al. (2012) used a phylogenetic analysis of a number of bacteria in order to propose cocktails of probiotic bacteria to reduce pathogens in food. Their approach is based on an investigation of a protein, MazF, which has an antimicrobial activity, and for which they propose an engineered variant. The phylogeny of the studied bacteria was instrumental in designing this new protein. The combination of molecular structure approaches with phylogenetic comparative analyses seems a promising venue to develop a variety of new molecules with desired properties.

Rich of its long history, the PCM seems to have a bright future both for addressing fundamental questions and for delivering applications.

References

Bartoszek K, Pienaar J, Mostad P, Andersson S, Hansen TF (2012) A phylogenetic comparative method for studying multivariate adaptation. J Theor Biol 314:204–215

Bininda-Emonds ORP, Cardillo M, Jones KE, MacPhee RDE, Beck RMD, Grenyer R, Price SA, Vos RA, Gittleman JL, Purvis A (2007) The delayed rise of present-day mammals. Nature 446:507–512

Blomberg SP, Garland T Jr, Ives AR (2003) Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution 57:717–745

Blomberg SP, Lefevre JG, Wells JA, Waterhouse M (2012) Independent contrasts and PGLS regression estimators are equivalent. Syst Biol 61:382–391

Bock KE (1966) The comparative method of anthropology. Comp Stud Soc Hist 8:269–280

Bock WJ (1959) Preadaptation and multiple evolutionary pathways. Evolution 13:194–211

Bock WJ (2003) Ecological aspects of the evolutionary processes. Zool Sci 20:279–289

Borgerhoff Mulder M, Nunn CL, Towner MC (2006) Cultural macroevolution and the transmission of traits. Evol Anthropol 15:52–64

Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis models and estimation procedures. Am J Hum Genet 19:233–257

Cheverud JM, Dow MM, Leutenegger W (1985) The quantitative assessment of phylogenetic constraints in comparative analyses: sexual dimorphism in body weight among primates. Evolution 39:1335–1351

Cliff AD, Ord JK (1981) Spatial and temporal analysis: autocorrelation in space and time. In: Wrigley EN, Bennett RJ (eds) Quantitative geography: a british view. Routledge & Kegan Paul, London, pp 104–110

Clutton-Brock TH, Harvey PH (1979) Comparison and adaptation. Proc R Soc Lond B 205:547–565

Cuvier G (1798) Tableau élémentaire de l'histoire naturelle des animaux. Baudouin, Paris

Darwin C (1859) On the origin of species by means of natural selection. John Murray, London

Eastman JM, Harmon LJ, Tank DC (2013) Congruification: support for time scaling large phylogenetic trees. Meth Ecol Evol 4:688–691

Felsenstein J (1973) Maximum-likelihood estimation of evolutionary trees from continuous characters. Am J Hum Genet 25:471–492

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

Felsenstein J (2004) Inferring phylogenies. Sinauer Associates, Sunderland

Felsenstein J (2005) Using the quantitative genetic threshold model for inferences between and within species. Phil Trans R Soc Lond B 360:1427–1434

Felsenstein J (2008) Comparative methods with sampling error and within-species variation: contrasts revisited and revised. Am Nat 171:713–725

Fisher RA (1930) The genetical theory of natural selection (a complete variorum edition, 1999). Oxford University Press, Oxford

Forster P, Toth A, Bandelt HJ (1998) Evolutionary network analysis of word lists: visualising the relationships between Alpine romance languages. J Quant Linguist 5:174–187

Freedman DA (2009) Statistical models: theory and practice (revised edition). Cambridge University Press, Cambridge

Garamszegi LZ, Møller AP (2010) Effects of sample size and intraspecific variation in phylogenetic comparative studies: a meta-analytic review. Biol Rev 85:797–805

Gittleman JL, Kot M (1990) Adaptation: statistics and a null model for estimating phylogenetic effects. Syst Zool 39:227–241

Grafen A (1989) The phylogenetic regression. Phil Trans R Soc Lond B 326:119–157

Grafen A, Ridley M (1997) A new model for discrete character evolution. J Theor Biol 184:7–14

Grandcolas P, Nattier R, Legendre F, Pellens R (2011) Mapping extrinsic traits such as extinction risks or modelled bioclimatic niches on phylogenies: does it make sense at all? Cladistics 27:181–185

Hadfield JD, Nakagawa S (2010) General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. J Evol Biol 23:494–508

Hadjipantelis PZ, Jones NS, Moriarty J, Springate DA, Knight CG (2013) Function-valued traits in evolution. J R Soc Interface 10(20121):032

Haeckel E (1887) Report on the Radiolaria collected by H.M.S. Challenger during the years 1873–1876. Her Majesty's Stationery Office, London

Hansen TF, Martins EP (1996) Translating between microevolutionary process and macroevolutionary patterns: the correlation structure of interspecific data. Evolution 50:1404–1417

Harvey PH, Pagel MD (1991) The comparative method in evolutionary biology. Oxford University Press, Oxford

Houwing-Duistermaat JJ, van Houwelingen HC, Terhell A (1998) Modelling the cause of dependency with application to filaria infection. Statist Med 17:2939–2954

Jetz W, Thomas GH, Joy JB, Hartmann K, Mooers AO (2012) The global diversity of birds in space and time. Nature 491:444–448

Jombart T, Pavoine S, Devillard S, Pontier D (2010) Putting phylogeny into the analysis of biological traits: a methodological approach. J Theor Biol 264:693–701

Kiekbaev DI (2003) Comparative law: method, science or educational discipline? Electronic journal of comparative law 7.3. url http://www.ejcl.org/73/art73-2.html

Kuhn TS, Mooers AO, Thomas GH (2011) A simple polytomy resolver for dated phylogenies. Meth Ecol Evol 2:427–436

Lamarck JB (1809) Philosophie zoologique. Flammarion (1994 edition), Paris

Lapiedra O, Sol D, Carranza S, Beaulieu JM (2013) Behavioural changes and the adaptive diversification of pigeons and doves. Proc R Soc Lond B 280(20122):893

Laurent G (1986) Cuvier et Lamarck: la querelle du catastrophisme. La Recherche 17:1510–1518

Legendre P (1993) Spatial autocorrelation: trouble or new paradigm? Ecology 74:1659–1673

Leroi AM, Rose MR, Lauder GV (1994) What does the comparative method reveal about adaptation? Am Nat 143:381–402

Lindenfors P, Jansson F, Sandberg M (2011) The cultural evolution of democracy: saltational changes in a political regime landscape. PLoS ONE 6:e28270

Losos JB (1994) An approach to the analysis of comparative data when a phylogeny is unavailable or incomplete. Syst Biol 43:117–123

Losos JB (2011) Convergence, adaptation, and constraint. Evolution 65:1827–1840

Lynch M (1991) Methods for the analysis of comparative data in evolutionary biology. Evolution 45:1065–1080

Mace R, Pagel M (1994) The comparative method in anthropology (with discussion). Curr Anthropol 35:549–564

Martins EP (1996) Conducting phylogenetic comparative studies when the phylogeny is not known. Evolution 50:12–22

Martins EP (2000) Adaptation and the comparative method. Trends Ecol Evol 15:296–299

Martins EP, Hansen TF (1997) Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. Am Nat 149:646–667 erratum vol 153, p 488

Münkemüller T, Lavergne S, Bzeznik B, Dray S, Jombart T, Schiffers K, Thuiller W (2012) How to measure and test phylogenetic signal. Meth Ecol Evol 3:743–756

Ollier S, Couteron P, Chessel D (2006) Orthonormal transform to decompose the variance of a life-history trait across a phylogenetic tree. Biometrics 62:471–477

O'Meara BC, Ané C, Sanderson MJ, Wainwright PC (2006) Testing for different rates of continuous trait evolution using likelihood. Evolution 60:922–933

Pagel M (1994) Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proc R Soc Lond B 255:37–45

Pagel M, Meade A (2006) Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. Am Nat 167:808–825

Pagel M, Meade A, Barker D (2004) Bayesian estimation of ancestral character states on phylogenies. Syst Biol 53:673–684

Pagel MD, Harvey PH (1988) Recent developments in the analysis of comparative data. Quart Rev Biol 63:413–440

Paradis E, Claude J (2002) Analysis of comparative data using generalized estimating equations. J Theor Biol 218:175–185

Pavoine S, Ollier S, Pontier D, Chessel D (2008) Testing for phylogenetic signal in phenotypic traits: new matrices of phylogenetic proximities. Theor Pop Biol 73:79–91

Pavoine S, Baguette M, Bonsall MB (2010) Decomposition of trait diversity among the nodes of a phylogenetic tree. Ecol Monogr 80:485–507

Pennell MW, Harmon LJ (2013) An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology. Ann NY Acad Sci 1289:90–105

Pepin KM, Lass S, Pulliam JRC, Read AF, Lloyd-Smith JO (2010) Identifying genetic markers of adaptation for surveillance of viral host jumps. Nat Rev Microbiol 8:802–813

Revell LJ (2009) Size-correction and principal components for interspecific comparative studies. Evolution 63:3258–3268

Rohlf FJ (2006) A comment on phylogenetic correction. Evolution 60:1509–1515

Santos JC, Cannatella DC (2011) Phenotypic integration emerges from aposematism and scale in poison frogs. Proc Natl Acad Sci USA 108:6175–6180

Shipley B (2013) The AIC model selection method applied to path analytic models compared using a d-separation test. Ecology 94:560–564

Simpson GG (1944) Tempo and mode in evolution. Columbia University Press, New York

Smith SA, Beaulieu JM, Stamatakis A, Donoghue MJ (2011) Understanding angiosperm diversification using small and large phylogenetic trees. Am J Bot 98:404–414

Spreitzer RJ, Peddi SR, Satagopan S (2005) Phylogenetic engineering at an interface between large and small subunits imparts land-plant kinetic properties to algal Rubisco. Proc Natl Acad Sci USA 102:17225–17230

Stone GN, Nee S, Felsenstein J (2011) Controlling for non-independence in comparative analysis of patterns across populations within species. Phil Trans R Soc Lond B 366:1410–1424

Thomas GH, Hartmann K, Jetz W, Joy JB, Mimoto A, Mooers AO (2013) PASTIS: an R package to facilitate phylogenetic assembly with soft taxonomic inferences. Meth Ecol Evol 4:1011–1017

Tiao GC, Reinsel GC, Xu DM, Pedrick JH, Zhu XD, Miller AJ, DeLuisi JJ, Mateer CL, Wuebbles DJ (1990) Effects of autocorrelation and temporal sampling schemes on estimates of trend and spatial correlation. J Geophys Res-Atmos 95:20507–20517

von Hardenberg A, Gonzalez-Voyer A (2013) Disentangling evolutionary cause-effect relation-
    ships with phylogenetic confirmatory path analysis. Evolution 67:378–387
Watt WB (2013) Causal mechanisms of evolution and the capacity for niche construction. Biol
    Philos 28:757–766
Whitney KD, Boussau B, Baack EJ, Garland T (2011) Drift and genome complexity revisited.
    PLoS Genet 7(e1002):092
Yan XH, Gurtler JB, Fratamico PM, Hu J, Juneja VK (2012) Phylogenetic identification of
    bacterial MazF toxin protein motifs among probiotic strains and foodborne pathogens and
    potential implications of engineered probiotic intervention in food. Cell Biosci 2:39

Chapter 2
Working with the Tree of Life in Comparative Studies: How to Build and Tailor Phylogenies to Interspecific Datasets

László Zsolt Garamszegi and Alejandro Gonzalez-Voyer

Abstract All comparative analyses rely on at least one phylogenetic hypothesis. However, the reconstruction of the evolutionary history of species is not the primary aim of these studies. In fact, it is rarely the case that a well-resolved, fully matching phylogeny is available for the interspecific trait data at hand. Therefore, phylogenetic information usually needs to be combined across various sources that often rely on different approaches and different markers for the phylogenetic reconstruction. Building hypotheses about the evolutionary history of species is a challenging task, as it requires knowledge about the underlying methodology and an ability to flexibly manipulate data in diverse formats. Although most practitioners are not experts in phylogenetics, the appropriate handling of phylogenetic information is crucial for making evolutionary inferences in a comparative study, because the results will be proportional to the underlying phylogeny. In this chapter, we provide an overview on how to interpret and combine phylogenetic information from different sources, and review the various tree-tailoring techniques by touching upon issues that are crucial for the understanding of other chapters in this book. We conclude that whichever method is used to generate trees, the phylogenetic hypotheses will always include some uncertainty that should be taken into account in a comparative study.

L. Z. Garamszegi (✉)
Department of Evolutionary Ecology, Estación Biológica de Doñana—CSIC,
Av. Américo Vespucio SN, 41092 Sevilla, Spain
e-mail: laszlo.garamszegi@ebd.csic.es

A. Gonzalez-Voyer
Conservation and Evolutionary Genetics Group, Estación Biológica de Doñana
(EBD–CSIC), Av. Américo Vespucio SN, 41092 Sevilla, Spain
e-mail: alejandro.gonzalez@ebd.csic.es

2.1 Introduction

According to evolutionary theory, all organisms evolve from a single common ancestor. Phylogenetic trees provide an elegant way to depict hypothesized ancestor–descendant relationships among groups of extant, or in some cases extinct, taxa, including all intermediate ancestors. In fact, the essence of all comparative methods lies in the varying degrees of shared ancestry among species that determine the expected similarity in phenotypes (Felsenstein 1985; Harvey and Pagel 1991). Given that phylogenies provide the necessary information about ancestor–descendant relationships, they are essential to any comparative analysis and each of them requires at least one phylogenetic hypothesis to be taken into account. Ultimately, the evolutionary conclusions will depend on the phylogeny used in the study.

Finding the true phylogenetic hypotheses from a large number of alternative trees is a very complex task. As the number of considered species increases, the number of potential phylogenetic resolutions also increases exponentially (Fig. 2.1). For practicing comparative biologists, questions about phylogenetic reconstruction are important to understand, because the constraints accumulated in this process should be considered in the next level of analysis, when the evolutionary inferences are being made. It is, therefore, necessary to have a good grasp of how phylogenies are estimated, what the assumptions and the main differences between the reconstruction methods are, and how the resulting trees can be tailored to a comparative study.

In this chapter, we provide a general overview on these steps and highlight that most reconstruction methods generate considerable uncertainty in the phylogenetic hypothesis. First, we define the essential terminology (see also Glossary at the end of the chapter), and then, we give a brief review of approaches that are most commonly used for phylogenetic reconstructions. Second, from the practical perspective, we explain how to obtain phylogenetic trees to match an interspecific data frame at hand and provide a guide for performing the most important tree-related exercises in a comparative study. Finally, we speculate about how the treatment of phylogenies (and the associated uncertainties they embed) will develop in the future. Although the issues that we present here might be obvious for most experienced users of the comparative methodology, who may skip this section, those who are new in this field may benefit from this discussion. Therefore, we recommend that beginners consult this chapter before continuing with the more advanced topics. Given that several primary resources are available that exhaustively review the phylogenetic reconstruction methods (Durbin et al. 1998; Ewens and Grant 2010; Hall 2004; Linder and Warnow 2006; Nei and Kumar 2000; Felsenstein 2004; Lemey et al. 2009; Page and Holmes 1998), here we only aim to provide a gentle introduction to the topic from the perspective of the readers of the book.
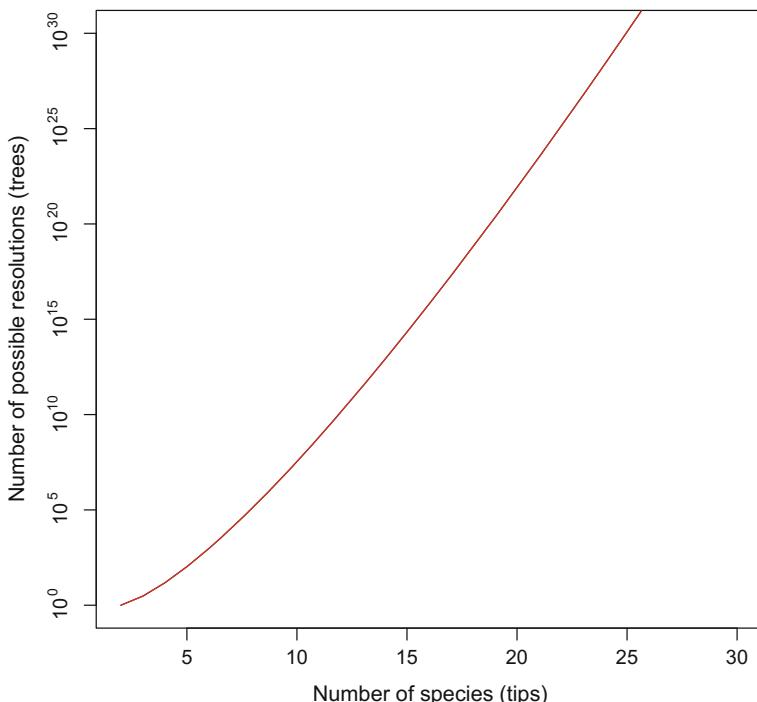
Fig. 2.1 The number of possible phylogenetic resolutions (trees) as a function of the number of species (tips)

2.2 Terminology

2.2.1 Homology and Homoplasy: Convergence and Divergence

Phylogenetic reconstruction methods are all based on the assumption that similarities in the traits (either morphological or genetic) used to estimate ancestor–descendant relationships are the result of homology, i.e., that they are similar because they were inherited from a common ancestor[1]. Only homologous traits can provide the necessary information about shared ancestry, in the form of shared similarities between species, and independent evolutionary history, in the form of differences between species the traits of interest. In the case of genetic sequences, differences among sequences in nucleotides (or amino acids) at specific positions are regarded as the result of divergence during the independent evolution following the speciation event. In a similar fashion, when morphological traits are used, the

---

[1] see also Glossary at the end of the chapter

differences in trait states are interpreted as resulting from independent evolution and thus are relevant for the clarification of the evolutionary relationships among species. Hence, homologous, gene sequences, or morphological characters presenting fewer differences are assumed to belong to species with shorter divergence times and thus more closely related than homologous traits with more differences. Similar selective regimes along independent branches of the phylogeny can result in both parallel and convergent evolution, which can in turn cause traits to present higher similarity than expected by chance. Such traits show homoplasy, which is not to be mistaken for homology. This is why phylogenetic reconstructions should rely on neutral markers, which are not under selection, or at the very least not strong or directional selection (Lemey et al. 2009; Page and Holmes 1998).

2.2.2 The Evolutionary History of Species as Reflected by a Phylogenetic Tree

The phylogenetic relationships among species are usually described in a tree format. The tree represents relationships among extant species at the tips (or leaves), but phylogenies may also include strains, higher taxonomic units, or even extinct taxa. In general, the taxa at the tips can also be termed operational taxonomic units (OTUs). The putative ancestors of the tips are represented by nodes at different levels (see Glossary for further definitions) that are connected to each other with branches. The number of nodes between two species is proportional to their evolutionary relationship, more closely related species are separated by fewer nodes than more distantly related species.

If a phylogenetic tree is rooted, one node is identified as the root that represents the most recent common ancestor of all the taxa, to which ultimately all other nodes descend through the links of branches. A rooted tree provides information about the sequence of evolutionary events that gave rise to the depicted relationships among the taxa, which allows defining ancestor–descendant relationships between nodes (those closer to the root are ancestral to those closer to the tips of the tree). On the contrary, unrooted trees do not provide information about ancestor–descendant relationships thus are not of interest for comparative analyses. Unrooted trees can lead to erroneous representations of expected similarity among taxa because sequences that are adjacent on an unrooted tree need not be evolutionarily closely related (Page and Holmes 1998).

Phylogenetic trees also generally include important information about the lengths of individual branches that connect the intermediate nodes and/or terminal tips and that can be used for inferences about evolutionary rates inherent to many phylogenetic comparative approaches. Branch lengths can represent the time that separates successive splits (divergence times) resulting in an ultrametric tree, or the number of evolutionary changes occurring in a molecular marker (e.g., nucleotide substitution) resulting in an additive tree. An important property of ultrametric trees is that all extant taxa have the same distance from the tip to the

root of the tree, or in other words the same distance separates any pair of extant taxa (but not extinct) in the tree when measured passing by the root (Page and Holmes 1998). The difference between additive and ultrametric trees has important consequences for comparative analyses (see Sect. 3.4). Therefore, users must be careful about the evolutionary assumptions associated with the different types of phylogenetic trees employed in comparative analyses.

2.2.3 Phylogenetic Uncertainty

Importantly, a phylogeny is only a hypothesis, thus it can always be replaced by a new one and some degree of uncertainty is always associated with it. There are two main types of phylogenetic uncertainty arising from the reconstruction itself.

First, there is uncertainty associated with the topology, i.e., with the degree to which the phylogeny represents true relationships between taxa. Because speciation events involve the splitting of one (ancestral) population into two incipient species, fully resolved phylogenies are, in theory, expected to be fully bifurcating (only two branches emerge from each node). Uncertainty in tree topology is, hence often reflected by the presence of polytomies, in which more than two descendant branches emerge from a single node. As the number of polytomies on a tree increases, so does the number of equally likely alternative phylogenetic hypotheses (one three-furcating phylogeny can be resolved in two ways). "Soft" polytomies reflect lack of sufficient data to adequately resolve the order of speciation events and thus the ancestor–descendant relationships on the tree. In that case, more information for different marker traits would be necessary to be able to resolve bifurcating relationships unambiguously. On the other hand, "hard" polytomies result from rapid or recent speciation events, when there has not been sufficient time for evolutionary changes to accumulate in the marker traits, which will then mask the order of the speciation events. Note that uncertainty in the tree topology can not only be reflected by polytomies, but also by estimates of support or robustness of the relationships among taxa (e.g., see bootstrapping or Bayesian posterior support below).

Second, there is also uncertainty associated with the branch lengths either reflecting time of divergence or the number of expected evolutionary changes. However, inferences from nucleotide substitutions or other measures of evolutionary change may sometimes be misleading because some events can be missed, for example, due to reversions to a state present in an ancestral sequence. Furthermore, the detected rate of evolutionary change in the phylogenetic marker can be affected by certain taxon-specific characteristics, for example, differences in body size, generation length, and metabolic rate, to name a few examples (Bromham 2011; Santos 2012). Different transformations exist to obtain branch lengths for a given topology (see Sect. 3.4).

How does uncertainty in the reconstruction of the phylogeny affect phylogenetic comparative analyses? Firstly, topological uncertainty can potentially

influence the estimates of the regression slope in analyses of associations between traits especially when the interspecific sample size is low. Inaccuracies in the topology have a stronger effect when alternative phylogenies involve species that are moved across the root (Blomberg et al. 2012). Changes closer to the tips, on the other hand, are less drastic in their effects (Martins and Housworth 2002; Symonds 2002). Uncertainty in species relationships at the tips of the tree might have a more important impact when assessing rates of phenotypic evolution, as topological errors could artificially inflate the estimates of interest if putative sister taxa present higher divergence than expected due to misplacement. Uncertainty in branch lengths can become much more problematic in analyses of rates of phenotypic evolution or in analyses of rates of diversification, as differences in branch lengths will directly affect parameter estimates.

There are several ways to incorporate phylogenetic uncertainty in comparative analyses, and most of these are discussed in details in other chapters (e.g., Chaps. 10–12). A simple method for controlling for uncertainty in tree topology involves repeating the analyses on each (or a subset of) alternative phylogenetic tree (Donoghue and Ackerly 1996). Furthermore, in analyses of correlations between traits or traits and environmental variables uncertainty in the branch lengths of the phylogeny (but not in the topology!) can be controlled by using parameter transformations (e.g., λ, α, and ρ) and maximum-likelihood or restricted maximum-likelihood methods which estimate the maximum-likelihood value of the parameter simultaneously with model fit (Martins and Hansen 1997; Pagel 1999; Freckleton et al. 2002). In general, uncertainties in the phylogenetic hypotheses can be effectively handled in the Bayesian (see Chap. 10) or in the Information Theoretic (see Chap. 12) statistical framework.

2.3 Assembling Phylogenies

2.3.1 Which Traits Are Appropriate for Phylogenetic Reconstruction?

For a trait to be used as a reliable phylogenetic marker, at least the following three criteria should be met: (i) similarities in trait values should be due to inheritance from a common ancestor, i.e., the trait should be homologous, (ii) the among-species variance in the trait should result from divergent evolution, and (iii) within-species variance is negligible compared to the among-species variance. The classical way of estimating relationships between species was to compare morphological characters (Linnaeus 1758), and taxonomy is still largely based on phenotypic characters. However, the increasing availability of molecular sequences and rapid development of a variety of analytical tools have led to the spread of genetic markers for phylogenetic reconstruction. Molecular data have an additional advantage over phenotypic characters, as they provide standard units comparable

across all living taxa. Given the overwhelming importance of molecular markers, in the rest of the discussion, we will limit ourselves to this focus with the note that most of the reviewed methodology works for morphological characters as well (as far as they fulfill the assumption of homology). We note, however, that although molecular markers are increasingly used for phylogenetic reconstruction and have virtually replaced morphological markers, this does not mean that phylogenetic inferences from gene sequences are necessarily free of uncertainty and/or of the problems posed by homoplasy.

2.3.1.1 Gene Trees Versus Species Trees

Different genetic mechanisms, such as gene duplications, genome reorganization, recombination, lateral gene transfer, have led to the diversity we observe today. Of all these sources of genetic variation, mutations (point mutations, insertions, and deletions) are used to infer relationships among genes. For the phylogenetic reconstruction to be reliable, the entire gene sequences being compared among taxa must have the same history. Recombination events, for example, are confounding because the recombining segments are not comparable. Recent gene duplication events leading to paralogous genes can also lead to unreliable phylogenetic reconstructions. Only the analysis of orthologous genes (homologous genes in different taxa that have started to evolve independently since divergence) provides information on the speciation events (Page and Holmes 1998). It is therefore important to ensure, a priori, that the genes employed in a phylogenetic analysis are orthologous to prevent flawed conclusions.

An important source of phylogenetic uncertainty is associated with the potential discrepancy between gene trees and species trees. Comparative analyses assume that the phylogeny represents the true and single evolutionary history of the species (or taxa) being analyzed. However, although intricately linked, the evolutionary history of genes can differ from that of the species, which leads to incongruence between the phylogenetic tree recovered for the gene and the unknown phylogenetic history of the species. Such differences can arise, for example, due to hybridization, gene duplication, horizontal gene transfer, and incomplete lineage sorting. The signatures that these processes leave on the gene trees can be utilized as phylogenetic signal to recover population parameters, evolutionary processes, and the species phylogeny itself (e.g., Nakhleh 2013; see also Chap. 3 in this book). Different approaches exist for inference of species trees. Some advocate the use of multiple genes (loci) concatenated into a single large matrix (supermatrix approach; Roquet et al. 2013) that can potentially reduce the effect of conflicting signal resulting from processes such as incomplete lineage sorting. Alternatively, others advocate the use of gene trees, where phylogenies are estimated independently for each locus and subsequently assembled into a large supertree (see Chap. 3).

2.3.1.2 Nuclear and Organelle DNA in Phylogenetic Reconstruction

An additional consideration for phylogenetic reconstruction is the choice of genome to be employed, as molecular phylogenies can be reconstructed using mitochondrial, chloroplast, or nuclear genes. Mitochondrial genes generally evolve faster and accumulate more substitutions than nuclear genes for various reasons (Galtier et al. 2009). For example, in *Drosophila*, mitochondrial genes have 4.5–9.0 times higher synonymous substitution rates (i.e., alterations in the nucleotide sequences do not affect the translated amino acid sequence) than average nuclear genes (Moriyama and Powell 1997). The chloroplast genome of plants, in contrast, presents a synonymous substitution rate which is on average 4 times lower than that of nuclear genes (Wolfe et al. 1989). Because of their faster rate of substitution, mitochondrial genes are generally more useful to resolve relationships among recently diverged species than nuclear genes, as the former are more likely to have accumulated the necessary substitutions. On the contrary, mitochondrial genes may be less informative regarding relationships dating further back in time because the phylogenetic signal is eroded. Such erosion of the signal results from the fact that only four bases constitute molecular sequences. Hence, as substitutions accumulate at a particular region of the gene, by mere chance, the probability increases that the nucleotide will change back to the base it had in the past and thus the difference with the ancestral sequence will be lost. This is referred to as saturation, for which models of sequence evolution attempt to correct. Furthermore, the mode of inheritance of the different genomes is also important, as in general the mitochondrial genome (and sometimes the chloroplast genome as well) is inherited from the maternal ancestor while the paternal copy is lost. For plants, where there is a higher frequency of hybridization, phylogenies reconstructed from chloroplast sequences might reveal only part of the story, as hybridization events would not be apparent. Given the differences in the rate of substitution between the two genomes, branch lengths could potentially differ between phylogenies reconstructed from organelle sequences compared with those reconstructed using nuclear sequences. To avoid such problems, recent attempts prefer to use a combination of genes from organelle and nuclear genomes.

2.3.2 From Nucleotide Sequences to Trees

2.3.2.1 Sequence Alignment

The first step of any phylogenetic reconstruction is to create a matrix containing the information on the states of marker traits in each species. In the case of phenotypic traits, the matrix will contain all traits aligned in columns with each trait coded as present or absent, or with different categories defining trait states. In the case of molecular data, the matrix is a sequence alignment, either involving protein sequences or nucleotide sequences. Some analyses involve both sequence

data and phenotypic traits. To obtain information on molecular sequences, one can use publicly available sources in GenBank in combination with the efficient search engines provided.

We must emphasize that obtaining sequence alignments is an error-prone process, and possibly one of the most challenging parts of the phylogeny reconstruction, as the raw GenBank data are unaligned and the processing of such data sometimes requires subjective decisions. Errors in the sequence alignment will carry through the entire process and be compounded, which can lead to incorrect phylogenetic reconstructions (and subsequent inferences about evolutionary mechanisms). Phylogenetic reconstruction methods are based on the assumption that compared traits or sequences are homologous (Page and Holmes 1998), and only correctly aligned sequences fulfill this assumption. A given sequence alignment is a hypothesis about homology of the nucleotides or amino acids and relies on the assumption that the sequences of different taxa have evolved from a common ancestral state. The aim of the sequence alignment is to position sequences along the matrix in such a way that homologous sites along the sequence are aligned in columns, as much as possible.

If two sequences have accumulated few substitutions, they will remain largely similar and the alignment will be straightforward. However, as sequences diverge and differences accumulate, it becomes increasingly difficult to make a sensible alignment. For two amino acid sequences with sequence identity below 25 %, finding a good alignment is highly challenging. Nucleotide sequences can be more problematic to align than amino acid sequences, since two random sequences of equal base composition will on average be 25 % identical merely due to chance. Therefore, when working with nucleotide sequences from coding genes, it is generally recommended to align such sequences at the amino acid level (Higgins and Lemey 2009), once issues about insertions and deletions in the nucleotide sequences are resolved (see below). Furthermore, some have advocated that together with the sequence identifiers the alignments themselves be made public when a new phylogenetic hypothesis is published.

Sequences do not always have the same length, thus gaps need to be inserted for the appropriate alignment. Gaps may represent either a deletion of one or more bases in a particular sequence, an insertion event—when one or more bases are incorporated into a sequence—or a combination of insertion and deletion events (insertion and deletion events are generally treated in the same manner). Repeats, when one or few nucleotides or an entire protein domain are repeated once or several times in sequence, can also cause problems. With short repeats inserted, it becomes highly difficult or virtually impossible to determine the correct alignment. Programs such as G-blocks (Castresana 2000) allow researchers to identify poorly aligned or highly divergent sections of the alignment, in which case the problematic section can be deleted in order to minimize errors. Insertion of gaps in an alignment is generally penalized (otherwise alignments with more gaps than nucleotides could result), while gaps at the end of a sequence are not penalized (as sequences might simply be missing sections at the end for various biological and experimental reasons). In some phylogenetic reconstruction programs (e.g.,

MrBayes Ronquist and Huelsenbeck 2003) gaps in the alignment are not informative, unless the user explicitly codes them as binary characters. Sequence data and sequence alignments are generally saved in text files, and the most commonly used formats are FASTA and NEXUS. A diversity of programs, many of them freely available online, for sequence visualization, alignment, and editing exist (e.g., MEGA, ClustalW, Mesquite, Seaview).

2.3.2.2 Models of Substitution

Assuming a proper alignment of nucleotide (or amino acid) sequences, the next step is to determine the best model of substitution, which provides a mathematical expression for the evolutionary transitions between states of the sequence data (e.g., through series of point mutations). Given that phylogenetic reconstructions often involve processes that unfold over potentially very long periods of time, it is simpler (and mathematically more tractable) to describe the models based on instantaneous probabilities. Hence, the models allow estimating the probability of observing any transition at a given point in time. Models may give different weights to different evolutionary transitions, for example, if some transitions are known to occur more frequently, these might be given a lower weight, and will have a lower impact on the reconstructed phylogeny, than rare transitions. For amino acids, substitution models are based on matrices that give different weights to all the possible transitions between the different amino acids based on knowledge about the frequency of transitions and similarity of biochemical properties (Higgins and Lemey 2009). For nucleotides, models weigh transitions (changes from a purine to a purine or from a pyrimidine to a pyrimidine) and transversions (changes from purine to pyrimidine or vice versa) differently. Substitution models also take base composition into account and estimate rate of molecular evolution. Furthermore, it is also possible to discriminate between processes that underline the occurrence of synonymous (i.e., not altering the composition of the translated protein) and non-synonymous (i.e., altering the amino acid sequence) mutations, and to correct for saturation. Models may be simple or complex, and the aim is to find the model which best describes the evolution of the data employed for phylogenetic reconstruction while at the same time minimizing the number of parameters that must be estimated. For the details of the particular models of sequence evolution, we refer to the primary literature (Durbin et al. 1998; Ewens and Grant 2010; Hall 2004; Linder and Warnow 2006; Nei and Kumar 2000).

Unfortunately, it is hard to decide a priory which model would be the most appropriate for the data (e.g., different mechanisms may apply to different taxa and markers), thus intuitively preferring one method over another might not be straightforward. Some care is warranted here, because the model chosen can have consequences for the outcome of tree reconstruction. Therefore, statistical methods may be needed, in which all potential models considered for sequence evolution are compared. Selecting from various models with different parameters is a model selection problem that practicing phylogeneticists must solve at the start of data

analysis based on some statistical means. This task is most commonly accomplished by comparing how different models incorporating particular scenarios for sequence evolution fit the data. Such model comparison strategy either follows a series of nested likelihood ratio tests or rely on Information Theoretic approaches based on Akaike Information Criterion (AIC-IT), for which statistical programs are largely available (e.g., ModelTest, Posada and Buckley 2004). Other approaches have also been developed, e.g., the one that applies decision theory to select models that minimize error in branch length estimation (Abdo et al. 2005; Minin et al. 2003), but different Bayesian methods have also generated noticeable popularity in molecular systematics (Alfaro and Huelsenbeck 2006; Arima and Tardella 2012). In some cases, different models can give similar results and issues about the uncertainties that are mediated by different tree estimation methods represent theoretical problems rather than manifest true concerns in practice.

2.3.2.3 Tree Reconstruction Methods

Once a substitution model has been chosen for the aligned sequences, several approaches are available for phylogenetic reconstruction (Fig. 2.2). We briefly review the most commonly adopted methods (Durbin et al. 1998; Ewens and Grant 2010; Hall 2004; Linder and Warnow 2006; Nei and Kumar 2000) and pinpoint how they contribute to our uncertainty about the evolutionary history of species. We name some computer programs that can perform such reconstructions. For practitioners interested in working with these approaches in the R statistical environment (R Development Core Team 2007), we recommend Paradis (2011).

Maximum Parsimony

Maximum parsimony is the method that relies on the fewest assumptions. It aims at finding the tree that involves the minimum number of evolutionary transitions in the marker trait. For example, imagine that we want to reconstruct the phylogenetic relationships between birds, rodents, and primates based on the presence of hair assuming a hairless ancestor. The most parsimonious phylogeny would yield that rodents are more closely related to primates than birds. This is because such a phylogeny would require only one evolutionary change (gain of hair in the common ancestor of primates and rodents), while a grouping of birds and primates together would necessitate two changes (gain of hair in two independent lineages).

Statistically, parsimony can be considered as a nonparametric method, it requires no parameters and does not estimate branch lengths (as the others below). Although it may appear simple, finding the most parsimonious resolution may be computer intensive for large number of species and long nucleotide sequences. Furthermore, it has received criticism, e.g., because the assumption about parsimony may be violated when evolution occurs at a rapid pace.
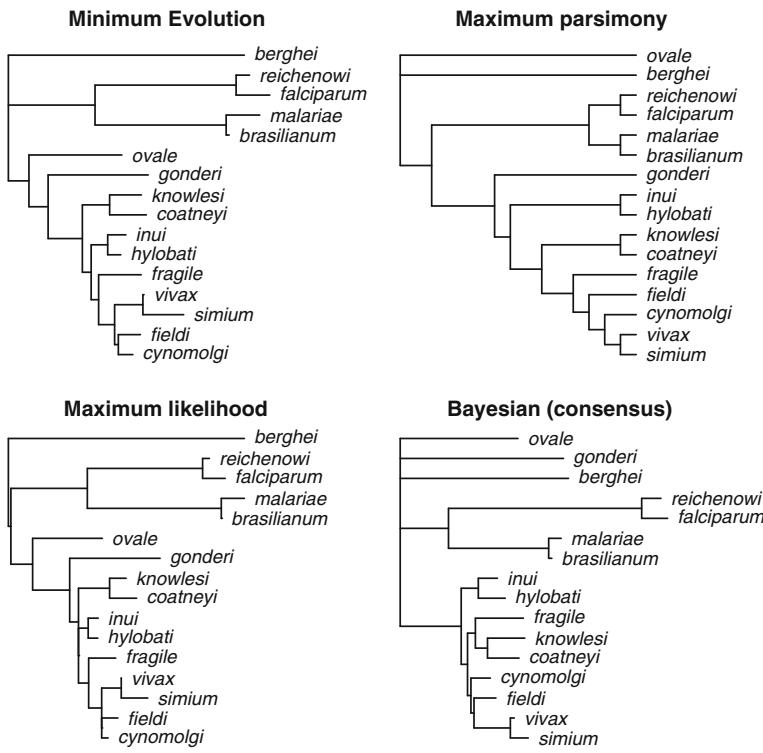
Fig. 2.2 Phylogenetic resolutions of 15 primate-infecting *plasmodium* (malaria) species, as revealed by the analysis of 18S rDNA sequences by different phylogeny estimation methods. The underlying data obtained form the GenBank and correspond to Leclerc et al. (2004)

Distance Methods

Several tree reconstruction approaches are distance methods that focus on the pairwise dissimilarities in nucleotide sequence between species. Such comparison of the genetic information across pairs of species results in a symmetric $N \times N$ matrix (N is the number of species), wherein each value defines the similarity/difference between two species based on a certain metric that considers a model of sequence evolution (as detailed above).

Once a distance matrix has been defined based on a given model of evolution, it can be used to describe the hierarchical (i.e., tree-shaped) associations among species, in which closely related species have higher similarity indexes than distantly related species. The difficulty is that one distance matrix mathematically defines more than one tree (whereas one tree defines only a single distance matrix), thus certain algorithms and evaluation methods are needed to find the most appropriate tree for a given matrix. These methods generally follow one of the two main strategies: either they aim at aggregating the most closely related species or splitting the most distantly related ones.

The most commonly used distance method is the *neighbor-joining method* (Saitou and Nei 1987), which aims at splitting the most distant observations by minimizing distances between the nearest neighbors (bottom–up clustering). The algorithm first builds a tree by connecting two randomly chosen species in a node and the remaining species in another node, then estimates total branch length. The combination of neighbors that results in the smallest length is retained and they are then removed from the distance matrix, which is then updated accordingly. This procedure is repeated until the tree becomes dichotomous. Further extensions to this basic method exist, which differ with respect to how they re-calculate the elements of the distance matrix after a split is retained. The advantage of the neighbor-joining method is that it is fast relative to other methods (e.g., maximum parsimony and maximum likelihood). However, it only gives a single tree as a result (i.e., does not incorporate uncertainty), which often depends on the model of evolution considered.

Another distance method is the *minimum evolution* method, which is based on an agglomerative process (Rzhetsky and Nei 1992). It assesses all possible topologies for a given distance matrix, and accepts the one that results in the smallest value for the sum of all branch lengths. The most common formula that can be used to estimate the sum of branch lengths from the distance matrix is based on ordinary least squares (OLS), which penalizes more for getting a long-branch wrong than for getting a short-branch wrong. Other methods also exist, and they differ in how they weight such differences. Given that the number of possible topologies dramatically increases with the number of species, it might be labor intensive for large datasets.

Maximum Likelihood

This approach is based on the estimation of a likelihood function that describes how a given tree is fitted to the observed sequence data. It considers an explicit model for character state evolution and a proposed phylogenetic tree with branch lengths. The degree to which a phylogeny explains the observed sequence data can be calculated as a likelihood, i.e., the conditional probability of the data (sequences) given the hypothesis (as defined by the evolutionary model and tree) considered. Finding the tree with the highest likelihood will tell us which phylogenetic hypothesis has the highest probability of producing the present-day sequences under the considered probabilistic model of sequence evolution. Maximum likelihood is known as a robust method, but evaluating likelihood surface can often require considerable time. Recently developed methods now allow for very rapid resolution of phylogenetic estimation even for datasets involving a large number of species, e.g., RAxML (Stamatakis 2006) or GARLI (Zwickl 2006). Importantly, likelihoods obtained for different trees are always conditional on an explicit model of evolution, which can be regarded either as a weakness or strength.

Bayesian Approaches

Bayesian methods for phylogenetic inference (Pagel et al. 2004b; Huelsenbeck et al. 2001) follow the Bayesian theorem to derive the posterior probability of a tree given the sequence data, as a function of the likelihood of the data given the tree and a prior belief in the general validity of the phylogenetic hypothesis (see more details on the Bayesian philosophy in Chap. 10). The posterior probability distribution of trees can be obtained via a Markov Chain Monte Carlo (MCMC) process, which proposes and evaluates a certain phylogenetic hypothesis along each state of the chain. These trees are obtained by the alteration of topologies, branch lengths or parameters of sequence evolution. A Metropolis–Hastings algorithm (Metropolis et al. 1953; Hastings 1970) is used to evaluate newly proposed trees, which will be accepted with a probability that is proportional to the ratio of their likelihood to that of the previous trees in the chain. If the chain is allowed to run enough along the universe of potential trees, the sampling will accumulate trees based on their likelihood. Then, the pool of sampled trees will form the posterior density of trees (i.e., the posterior density of topologies, branch lengths, and parameters of the model of evolution) in proportion to their frequency of occurrence. The resulting Markov sample will consist of thousands or millions of trees, with each of them being represented according to how they fit the genetic data. Therefore, in contrast to other methods that provide a single tree as solution, the Bayesian approach has the capacity to capture the uncertainty in the phylogenetic hypothesis in the form of the distribution of similarly likely trees. The often-emphasized shortcoming of this method is that it requires prior information on the topologies, branch lengths, and other parameters of the model of substitution, which is often challenging.

2.3.2.4 Tree Evaluation

Tree reconstruction from molecular data does not result in an unambiguous translation, but each resulting tree can be described by a certain degree of uncertainty that underlies the derived branching pattern. Different methods are known that quantitatively test the reliability of an inferred tree, and these provide support values for the topology (not the branch length!) of the tree that are presented at the nodes of the phylogeny. These values assess the confidence of the particular nodes with lower values, suggesting higher uncertainty associated with the node.

Bootstrapping methods can be generally applied to most phylogenetic estimation strategies (such as the neighbor-joining, minimum evolution, and maximum-likelihood methods). The bootstrap approach implements a resampling iteration, in which from m number of aligned sequences consisting of n number of nucleotides, n number of nucleotides are chosen randomly with replacement until constituting a new set of m sequences. Then based on this bootstrap sample, the tree is reconstructed by using exactly the same reconstruction method as was used for the

original alignments, and the topology of the two trees is compared. In this comparison, each node on the original tree that is identical with the analog node of the bootstrap tree receives a score of 1, while differing nodes are scored as 0. This procedure is repeated several hundred or thousand times, through which the probability of exact matches between the nodes of the bootstrap and that of the original tree (i.e., the percentage of times of scores 1) is recorded. These bootstrap values can be used for making inferences about the reliability of the original tree, with values above 95 % or higher, suggesting reliable topology.

Trees obtained through a Bayesian process, which already provides a pool of a large number of trees in the posterior sample can be summarized by Bayesian posterior support values. These are simply the proportion of trees in the posterior sample in which the node/clade is present.

2.3.3 Supertrees

Recent days' practice rarely requires the above exercise with nucleotide sequences. Instead, practitioners can rely on publically available supertrees that summarize the accumulated phylogenetic information for large taxonomic groups, such as mammals or birds (Ahlquist 1990), in an electronic format (e.g., Arnold et al. 2010; Jetz et al. 2012a; Bininda-Emonds et al. 2007). Supertrees offer a huge practical benefit for users, as they can ideally upload the list of species then obtain a fully matching phylogenetic tree in one click. Uncertainties about the phylogenetic hypothesis can be stored in series of trees representing alternative resolutions and branch lengths that can be taken forward to the next level of analysis. However, the creation of supertrees has its own caveats, and these should be taken into account when these resources are exploited. More details about the reconstruction and the use of supertrees can be found in Chap. 3 in this book.

2.3.4 The Classical Way of Assembling Trees by Hand

Historically, and for completeness, we need to mention that before the spread of nucleotide sequences and supertrees, the common practice for obtaining phylogenies was based on a tailoring exercise that was accomplished by hand. Practitioners of this method accumulated big piles of hard copies of papers that presented phylogenetic information for their taxon of interest (e.g., fishes, birds, and mammals). Then for a particular comparative study, they went through this collection and looked for phylogenetic information for the species in the given dataset. To combine this information, a backbone phylogeny of families or other higher taxonomic groups (e.g., by using large tapestry trees like Sibley and Ahlquist 1990) was created, on which each species was subsequently added if appropriate resolutions were available in the phylogenetic literature. When

judgments about phylogenetic associations were based on different studies using different methods and/or markers, the combination of branch lengths across sources was impossible. Given that the whole process was done by hand, in most cases handling a large number of alternative phylogenies was impractical and cognitively challenging. Although this tailoring exercise has lost its importance, it may be still useful when the effect of alternative resolutions of a particular species or clade is investigated.

2.4 Manipulating Phylogenies for Comparative Analyses

In the course of a comparative study, the investigator is typically required to work directly with the phylogenetic tree (or trees) before entering it into a statistical model. Although we cannot be exhaustive, below we highlight the most common tasks that emerge in an average phylogenetic comparative project. In Table 2.1, we list the most common tree manipulation exercises, for which we also provide working examples relying on the R statistical environment (R Development Core Team 2007) in the Online Practical Material (hereafter OPM, available at http://www.mpcm-evolution.com). For a more comprehensive list of software for phylogenetic reconstruction and manipulation, see http://evolution.genetics.washington.edu/phylip/software.

The first step of a comparative analysis is to import the phylogenetic tree. As different packages can be used for tree manipulation (and analysis), it is now inevitable to work with tree formats that are generally readable in various platforms. The most common tree formats that can be flexibly exported and imported are the Newick or NEXUS formats, which define the branching patterns (topology and branch lengths) by using a standard text code (see Chap. 4). These codes also allow importing multiple trees with additional information, such as bootstrap or Bayesian support of nodes or character state probabilities, for example, from ancestral state reconstruction. Given the increasing popularity of phylogenetic generalized least squares (PGLS) methods that relies on the expected variance-covariance matrix of species (see Chap. 5 for more details), it has also become common to transfer phylogenetic information in the form of a variance-covariance matrix.

The second step of any analysis is to ensure the taxa represented in the phylogenetic tree match the species in the database. It is very rarely the case that all species in the database will be present in the tree and vice versa, and sometimes a few annoying typos can cause incompatibilities. Moreover, some approaches even require the list of species to be in the same order as on the phylogenetic tree (or in the corresponding variance-covariance matrix). These tasks may seem simple and obvious, but our experience is that, if not done automatically by the program or if an error message does not appear, most students fail to check the correspondence between the interspecific data and phylogeny. Entering unmatched data and tree into the analysis may result in a situation where a random phylogenetic tree is used

Table 2.1 Most common tree operation tasks that emerge in comparative phylogenetic studies. This list serves merely an illustration purpose and does not intend to provide an exhaustive summary of applications

Task	Purpose	Software
Creating a phylogenetic tree with random resolution, or a star phylogeny for a list of species	Starting tree for a manual tree-building exercise	Mesquite
	Null hypothesis generation for testing for phylogenetic signal	R packages: *ape* *phytools*
Adding/removing species	Pruning tree for an interspecific dataset	Mesquite TreeView TreeEdit
	Testing the effect of particular clade on the phylogeny	R packages: *ape* *phytools*
Moving branches	Building a tree manually	Mesquite TreeView TreeEdit
Altering branch lengths	Testing the effect of different evolutionary models and the mode of trait evolution	R packages: *geiger* *phytools*
Creating an ultrametric tree from non-ultrametric tree	Fulfilling assumptions of approaches that require ultrametric tree	R packages: *ape*
Rooting an unrooted tree	Fulfilling assumptions of approaches that require rooted tree	Mesquite TreeEdit R packages: *ape*

(continued)

Table 2.1 (continued)

Task	Purpose	Software
Creating alternative resolutions	Testing for the effect of an alternative phylogenetic hypothesis	Mesquite R packages: *ape*
Resolving/creating polytomies	Building a tree manually Testing for the effect of an alternative phylogenetic hypothesis Fulfilling assumptions of approaches that require fully dichotomous tree	Mesquite TreeView TreeEdit R packages: *ape*
Comparing tip labels with the list of species in the dataset	Verifying one to one match between the phylogeny and the data Fulfilling assumptions of approaches that require that the list of species in the phylogeny matrix is the same as in the dataset	Excel R packages: *geiger*
Tree visualization (adding node labels, branch lengths, ladderizing…)	Interpretation and publication	MacClade Figtree Mesquite TreeView TreeEdit R packages: *ape* *phytools*
Plotting trait values on the phylogeny	Interpretation and publication	Mesquite MacClade R packages: *ape* *phytools*

(continued)

Table 2.1 (continued)

Task	Purpose	Software
Handling a large number of trees	Accounting for phylogenetic uncertainty	Mesquite MrBayes R packages: *ape*
Saving/exporting trees in different formats	Importing phylogenies into different softwares	MacClade Figtree Mesquite TreeView TreeEdit R packages: *ape* *phytools*
Simulating trees	Phylogenetic randomization	Compare MacClade Mesquite R packages: *ape* *phytools*
Calculating variance-covariance matrix from tress	Implementing phylogenetic information into the PGLS framework	R packages: *ape*

in the comparative tests, thus the essence of the entire study may have been lost. Therefore, we recommend students to incorporate as standard practice the comparison of species lists in the database and on the phylogeny prior running the analyses.

Once a tree is imported and matched with the list of species, it might be of interest to verify how the phylogeny looks. We note that the graphical representation of phylogenies is the most comprehensible way for a human observer to interpret the phylogenetic associations between species (note that none of the phylogenetic approaches use the trees as we do), thus it is important to visualize trees to check for potential errors and also to derive evolutionary inferences. We present some basic tree visualization methods in the OPM. Chapter 4 gives an extensive list of solutions for generating more enhanced graphical representations of the phylogeny and results of some comparative analyses that can be used for biological interpretations. Notably, character states of both extant species and ancestral nodes, bootstrap support or Bayesian posterior probabilities, geographic projections, and other components of the comparative results can also be plotted on the phylogenies that further enhance interpretations.

The user may sometimes want to modify the branch lengths of the tree either to fit a particular evolutionary model requiring specific branch lengths, or to obtain an ultrametric tree from an additive tree. The latter task is quite important as most comparative analyses assume that the tree is ultrametric, as the majority of analyses deal with evolution of phenotypic traits of extant species with the underlying assumption is that the time available for phenotypic evolution is the same for all taxa. However, additive trees will violate this assumption, and in some programs, it is still possible to run models with trees that conflict the assumption about ultrametricity without any warning (e.g., the commonly used trees with unit branch length are not ultrametric). Additive trees can be used when the taxon sampling times differ, and these are expected to in turn impact the evolution of the trait(s) of interest, for example, when dealing with viruses, or if the investigator has some a priori assumption that the molecular rate of evolution of the genes used to reconstruct the phylogeny directly impacts on the rate of phenotypic evolution (of course by avoiding circularity). It is always preferable to employ trees that have been calibrated to reflect time based on information such as fossils and geological events, however if such trees are not available an option is to use nonparametric rate smoothing (Sanderson 1997) to transform the tree. Other methods also exist to transform branch lengths of a given tree or to estimate branch lengths for a given topology. A common transformation is to simply apply unit length to all branches, i.e., equal branch lengths. Grafen (1989) proposed another transformation that involves first assigning a height to each node of the tree of one less than the number of species below the node, then branch lengths are the difference between the height of the upper and lower nodes, resulting in an ultrametric tree. Other methods transform branch lengths of a tree based on specific models of trait evolution such as Brownian motion (Freckleton et al. 2002; Pagel 1999), the Ornstein–Uhlenbeck processes (Hansen 1997; Martins and Hansen 1997), or different rates of evolution toward the root than toward the tips of the tree (Grafen 1989; Pagel 1999), to name a

few. Users must be aware of the evolutionary assumptions associated with branch length transformations, whether these fit the traits and model they are studying and ensure that they do not violate assumptions of the comparative methods they wish to employ.

Another frequently applied exercise with phylogenetic trees is the resolution of polytomies, as some comparative methods require fully bifurcating trees. One way to handle such a situation is to rearrange the polytomy into an aleatory set of bifurcations separated by zero-length branches. The resolution, in practice, is virtually the same as the polytomy, as the distance between taxa will remain the same, but additional nodes (putative ancestors) are added to the tree to resolve the order of splitting events randomly. Importantly, polytomies represent uncertainties about the topology, thus all possible alternative resolutions that can be produced by random dichotomization should ideally evaluated in the comparative models to test their effects on the results. A specific (extreme) case of polytomy is when all species are connected into a single node (root) with the same branch length. Such star phylogeny can also be used for fitting evolutionary models leading to results that will be equivalent with what could be obtained without controlling for phylogeny (the benefit of fitting a model with a star phylogeny is that this model will have the same number of estimated parameters than the model relying on the true phylogeny, and this property can be exploited for model comparison).

If a species is not present on the tree, but the investigator has information, either from taxonomy or other phylogenetic reconstructions, about the possible placement of the species it can be added to the tree. If its phylogenetic relations are fully known, the new species can be added onto the tree with complete bifurcation, but if there is uncertainty about the exact resolution, it can be lumped within a polytomy. Similarly, if phylogenetic information is accumulating, one can also update the tree by moving specific branches (without adding new tips). Once an alternative resolution is acquired, it might be warranted to run sensitivity analyses to determine the influence of tip additions or branch movements on the results.

Owing to the recent spread of phylogenetic simulations (see Chap. 13), it is becoming more and more routine to work with simulated trees, which can be created with a relative ease even under different evolutionary scenarios. For example, the investigator might wish to contrast the comparative results obtained from the original tree with those that correspond to a tree that was simulated under certain evolutionary conditions, or to a large number of randomly simulated trees that serve as a null hypothesis. Moreover, simulated trees are frequently used in simulation studies that test for the performance of particular comparative approaches (e.g., Revell and Reynolds 2012; de Villemereuil et al. 2012; FitzJohn et al. 2009b; Ives et al. 2007).

2.5 Discussion

2.5.1 Importance of Incorporating Phylogenetic Uncertainty in Comparative Analyses

A key issue that repeatedly emerged in this chapter is that all phylogenetic reconstructions inescapably involve some uncertainty, which is manifested as alternative hypotheses about the topology of the tree and branch lengths. This uncertainty is inevitable, because we are attempting to reconstruct processes having occurred in the very distant past, for which data are hardly available (see Chap. 22). The reconstruction of the evolutionary ancestor–descendent relationships among taxa based on any trait is a complex task that involves several steps (as we reviewed above), each of them encompassing uncertainty in the result.

The reconstruction methods assume that similarities are the result of shared ancestry, but evolutionary processes can deviate from the assumed patterns of inheritance. For example, horizontal gene transfer or hybridization can lead to contradictory phylogenetic signals. Furthermore, rapid speciation can result in incomplete lineage sorting, where ancestral polymorphisms are not fully resolved prior to second speciation events. Another source of uncertainty is associated with the extent to which the morphological or genetic marker trait represents the historical process of diversification of species. Uncertainties also arise due to the alignment, missing data, differences between models of substitution and the parameterization of the models. Finally, uncertainty in the branch lengths can also arise during estimation of divergence times. It is important to be aware of the different types of uncertainty and that it can be due to a diversity of processes. Rather than ignoring it, we advocate that a straightforward scientific approach should attempt as much as possible to incorporate this uncertainty into comparative analyses and assess its effects on the results. Ultimately, our understanding about how nature operates based on any estimation process from empirical data is unavoidably loaded with certain degree of uncertainty, which biologists should appreciate.

2.5.2 Future Directions

2.5.2.1 Increasing Amount of Information

The exponential increase in the availability of sequence information in public databases (e.g., GenBank has sequences for nearly 260,000 described species; Benson et al. 2011) as well as the number of different species for which a complete genome sequence becomes available are likely to have an important impact on both phylogenetic reconstruction and comparative biology. Along this progress, one challenge will be to determine which genes are most reliable to reflect the evolutionary history of species. Suitable markers are generally expected to improve the

"signal-to-noise" ratio, i.e., the amount of true phylogenetic information with respect to the embedded uncertainty. Fast-evolving genes will be good candidates to reconstruct recent events, while conserved markers will be more useful for deep relationships (but see Kälersjö et al. 1999). Therefore, the combination of markers of the two types may be fruitfully exploited for phylogenetic reconstructions. Ideally, markers should present low variation in copy number across taxa in order to also be of use to estimate within-species variation (Wu et al. 2013).

An important characteristic is universality, especially when attempting to reconstruct relationships among very distant taxa. The use of universal markers will likely play an important role in defining relationships at the root of the "tree of life" (Burleigh et al. 2011; Desluc et al. 2005). The growing consensus in deep phylogenetic relationships will allow creating a backbone constraint tree to define monophyletic groups for the reconstruction of megaphylogenies. Different approaches for the reconstruction of such megaphylogenies have already been proposed (Jetz et al. 2012b; Roquet et al. 2013; Thomas et al. 2013; Bininda-Emonds et al. 1999), and we expect that in the future it will become increasingly common for comparative biologists to employ such methods to reconstruct phylogenies specifically tailored to suit their needs. The reconstruction of megaphylogenies based either on supermatrices or supertrees will also be of interest in that it may generate standard phylogenies that can be used in different studies. However, availability of such standard trees should not lead to a biased perception of certainty in the reconstruction.

2.5.2.2 Improving Comparative Methodologies

How can phylogenetic comparative methods incorporate uncertainty in the phylogenetic reconstruction? Bayesian methods present a convenient means of incorporating both uncertainty in the phylogeny as well as uncertainty in parameter estimates in a single analysis (Huelsenbeck et al. 2001). Methods are currently available allowing researchers to undertake analyses using, for example, a subsample of phylogenies from the posterior distribution of trees from a reconstruction using Bayesian methods (Amcoff et al. 2013; Gonzalez-Voyer et al. 2008; Pagel and Meade 2006; Santos-Gally et al. 2013; Pagel et al. 2004a; de Villemereuil et al. 2012). An alternative, yet unexplored, approach based on Information Theory and model comparison is presented in Chap. 12 of this book. Different alternative methods have also been developed to incorporate uncertainty due to incompletely sampled phylogenies in analyses of rate of diversification and trait-dependent speciation or extinction (FitzJohn et al. 2009a; Morlon et al. 2011).

In the current state of the art, empirical studies are needed to determine the influence of phylogenetic uncertainty on comparative results in relation to various evolutionary questions. The importance of the consideration of phylogenetic uncertainty in comparative studies is well founded on theoretical bases, but we lack empirical data on how much alternative phylogenies can affect the comparative findings in general. In addition, simulation studies will no doubt play an

important role in determining to what extent phylogenetic uncertainty can influence the results of comparative analyses and whether certain methods are more vulnerable to uncertainty in the topology or branch lengths. Such a simulation may target questions regarding the importance of uncertainties accumulated in certain nodes or branches of the phylogenetic tree (see Blomberg et al. 2012; Martins and Housworth 2002; Symonds 2002 for example).

2.5.2.3 Evolutionary Processes Not Represented in a Tree

Although phylogenetic trees provide useful and simple means of representing the evolutionary history of a group of taxa, a phylogenetic tree does not represent evolutionary processes that deviate from the assumption of homology but still contribute to the diversification of species. In particular, horizontal inheritance of traits disrupts the tree-shaped representation of evolutionary processes that can only cope with vertical events. Mechanisms that play an important role in generating genetic variability across populations and breeds such as gene flow and very recent fluctuations in population size can have profound effects on genetic diversity of populations and expected similarities, which will bias estimates of phylogenies (Kalinowski 2009). Cultural evolution is another example, in which horizontal transmission of information plays an important role in shaping the interspecific variance of phenotypes. Accordingly, the evolutionary history of populations or breeds does not necessarily follow the hierarchical, bifurcating structure of phylogenetic trees, but relationships between taxa may be better described by networks that allow incorporating horizontal processes of transmission (i.e., reticulation). Developments of comparative analyses that can account for such a network structure delineate a fascinating research direction (Stone et al. 2011). The increasing availability of next generation sequencing applied to sample genome-wide polymorphisms across many populations will likely provide the necessary marker traits for methods that can reconstruct both horizontal and vertical events. The challenge will lie in developing the methods to adequately represent the evolutionary histories.

Glossary

Additive tree/ phylogeny	A phylogeny is termed additive when the tips are not all equidistant from the root. In an additive phylogeny branch lengths represent the number of expected substitutions, therefore differences among taxa in the rate of molecular evolution will lead to differences in branch lengths.
Branch	A continuous line that connects two nodes or a node to a tip in the phylogeny.

Branch length	Represents the "distance" between the two nodes or the node and tip connected by the branch. The "distance" can be measured in number of evolutionary transitions (if the phylogeny is reconstructed using maximum parsimony methods), number of expected substitutions, which is an estimate of the rate of molecular evolution, or divergence times.
Gene duplication	When a second copy of an existing gene emerges within a single genome. Gene duplication is a major mechanism by which new genetic material is generated.
Homology	Shared similarity between taxa that is due to inheritance from a common ancestor.
Homoplasy	Similarity between taxa that results from convergent evolution, for example due to similar selection pressures.
Horizontal gene transfer	The transfer of genetic material between individuals of different species, and which is not the result of inheritance from a common ancestor.
Hybridization	Mating between individuals of two distinct species of plants or animals resulting in viable offspring.
Incomplete lineage sorting	Occurs when coalescence times of alleles are within the time span of speciation events or shorter. Incomplete lineage sorting results in gene genealogies that are not concordant with the species phylogeny.
Nodes	Represent the putative ancestors of the taxa represented in the phylogeny.
Orthologous genes	Genes originating from a common ancestor (i.e. homologous genes) that have undergone independent evolution following a speciation event.
Parallel or convergent evolution	Evolution of phenotypes or sequences under similar selective regimes leading to higher similarities than would be expected based on the degree of shared ancestry.
Paralogous genes	Genes originating from a duplication event recent enough to reveal their common ancestry.
Polytomy	When more than two branches originate from a single node in the phylogeny. Polytomies reflect uncertainty in the timing of speciation events, either because of lack of sufficient data to determine the order of events

	with confidence (so called "soft polytomies") or because the speciation events were so rapid there was insufficient time for the necessary substitutions to discriminate between the timings of the speciation events to accumulate (so called "hard polytomies").
Root	Represents the most recent common ancestor of all the tips (taxa) in the phylogeny. All branches of the phylogeny lead to the root and the root connects all nodes.
Saturation	Occurs when two aligned, presumably orthologous, sequences have accumulated such an elevated number of repeated substitutions that these provide a poor estimate of their time of divergence. Saturation occurs because there is a higher probability of reverse mutations (changes to a nucleotide present in the past) as time of divergence increases and hence apparent differences between orthologous sequences become lower than expected based on the time of divergence.
Substitution rate	Also referred to as molecular evolution rate, it is the rate at which organisms accumulate genetic differences over time, it is usually calculated as the number of substitutions per site per unit time. Non-synonymous and synonymous substitutions can be discriminated depending on whether changes in the nucleotide sequence affect the translated amino acide sequences or not, respectively.
Tips	Also called leaves (following the tree analogy for phylogenies) they are the taxa whose relationships are being estimated with the phylogeny
Ultrametric tree/ phylogeny	A phylogeny is termed ultrametric when all the tips are equidistant from the root. In other words the distance between any two species in the tree is the same as long as the path crosses the root of the tree. In ultrametric trees the branch lengths usually represent divergence times. Ultrametric trees can also be estimated under the assumption of a constant rate of substitution that is the same for all taxa, also called a molecular clock. However, recent studies with diverse species have called into question the molecular clock showing that the rate of molecular evolution varies among even closely related species and is correlated with species-specific traits and even environmental variables.

References

Abdo Z, Minin VN, Joyce P, Sullivan J (2005) Accounting for uncertainty in the tree topology has little effect on the decision-theoretic approach to model selection in phylogeny estimation. Mol Biol Evol 22 (3):691–703. doi:10.1093/molbev/msi050

Alfaro ME, Huelsenbeck JP (2006) Comparative performance of Bayesian and AIC-based measures of phylogenetic model uncertainty. Syst Biol 55(1):89–96. doi:10.1080/10635150500433565

Amcoff M, Gonzalez-Voyer A, Kolm N (2013) Evolution of egg dummies in tanganyikan cichid fishes: the roles of parental care and sexual selection. J Evol Biol 26:2369–2382. doi:10.1111/jeb.12231

Arima S, Tardella L (2012) Improved harmonic mean estimator for phylogenetic model evidence. J Comput Biol 19(4):418–438. doi:10.1089/cmb.2010.0139

Arnold C, Matthews LJ, Nunn CL (2010) The 10k Trees website: a new online resource for primate phylogeny. Evol Anthropol 19:114–118

Benson DA, al. e (2011) GenBank. Nucleic Acids Res 39:D32–D37

Bininda-Emonds O, Gittleman JL, Purvis A (1999) Building large trees by combining phylogenetic information: a complete phylogeny of the extant Carnivora (Mammalia). Biol Rev 74:143–175

Bininda-Emonds ORP, Cardillo M, Jones KE, R DEM, Beck RMD, Grenyer R, Price SA, Vos RA, Gittleman JL, Purvis A (2007) The delayed rise of present-day mammals. Nature 446:507–512

Blomberg SP, Lefevre JG, Wells JA, Waterhouse M (2012) Independent contrasts and PGLS regression estimators are equivalent. Syst Biol 61(3):382–391. doi:10.1093/sysbio/syr118

Bromham L (2011) The genome as a life-history character: why rate of molecular evolution varies between mammal species. Phil Trans R Soc B 366:2503–2513. doi:10.1098/rstb.2011.0014

Burleigh JG, Bansal MS, Eulenstein O, Hartmann S, Wehe A, Vision TJ (2011) Genome-scale phylogenetics: inferring the plant tree of life from 18,896 gene trees. Syst Biol 60(2):117–125. doi:10.1093/sysbio/syq072

Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540–552

de Villemereuil P, Wells JA, Edwards RD, Blomberg SP (2012) Bayesian models for comparative analysis integrating phylogenetic uncertainty. BMC Evol Biol 12. doi:10.1186/1471-2148-12-102

Desluc F, Brinkmann H, Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. Nature Rev Genet 6(5):361–375

Donoghue MJ, Ackerly DD (1996) Phylogenetic uncertainties and sensitivity analyses in comparative biology. Phil Trans R Soc B 351:1241–2149

Durbin R, Eddy S, Krogh A, Mitchison G (1998) Biological sequence analysis. Cambridge University Press, Cambridge

Ewens WJ, Grant GR (2010) Statistical methods in bioinformatics: an introduction. Springer Science and Business Media, New York

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

Felsenstein J (2004) Inferring phylogenies. Sunderland, Sinauer Associates

FitzJohn RG, Maddison WP, Otto SP (2009a) Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. Syst Biol 58(6):595–611. doi:10.1093/sysbio/syp067

FitzJohn RG, Maddison WP, Otto SP (2009b) Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. Syst Biol 58:595–611

Freckleton RP, Harvey PH, Pagel M (2002) Phylogenetic analysis and comparative data: a test and review of evidence. Am Nat 160(6):712–726. doi:10.1086/343873

Galtier N, Jobson RW, Nabholz B, Glemin S, Blier PU (2009) Mitochondrial whims: metabolic rate, longevity and the rate of molecular evolution. Biol Lett 5 (3):413–416. doi:rsbl.2008.0662 [pii] 10.1098/rsbl.2008.0662

Gonzalez-Voyer A, Fitzpatrick JL, Kolm N (2008) Sexual selection determines parental care patterns in cichlid fishes. Evolution 62 (8):2015–2026. doi:EVO426 [pii] 10.1111/j.1558-5646.2008.00426.x

Grafen A (1989) The phylogenetic regression. Phil Trans R Soc B 326(1223):119–157

Hall BG (2004) Phylogenetic trees made easy: a how-to manual. Sinauer Associates Inc, Sunderland

Hansen TF (1997) Stabilizing selection and the comparative analysis of adaptation. Evolution 51(5):1341–1351

Harvey PH, Pagel MD (1991) The comparative method in evolutionary biology. Oxford University Press, Oxford

Hastings WK (1970) Monte carlo sampling methods using markov chains and their applications. Biometrika 57(1):97–109. doi:10.2307/2334940

Higgins D, Lemey P (2009) Multiple sequence alignment. In: Lemey P, Salemi M, Vandamme A-M (eds) The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing. Cambridge University Press, Cambridge, pp 68–96

Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310–2314

Ives AR, Midford PE, Garland T (2007) Within-species variation and measurement error in phylogenetic comparative methods. Syst Biol 56(2):252–270. doi:10.1080/10635150701313830

Jetz W, Thomas GH, Joy JB, Hartmann K, Mooers AO (2012a) The global diversity of birds in space and time. Nature 491(7424):444–448. doi:10.1038/nature11631

Jetz W, Thomas GH, Joy JB, Hartmann K, Mooers AO (2012b) The global diversity of birds in space and time. Nature 491:444–448. doi:10.1038/nature11631

Kälersjö M, Albert VA, Farris JS (1999) Homoplasy increases phylogenetic structure. Cladistics 15(1):91–93. doi:10.1111/j.1096-0031.1999.tb00400.x

Kalinowski ST (2009) How well do evolutionary trees describe genetic relationships among populations? Heredity 102:506–513. doi:10.1038/hdy.2008.136

Leclerc MC, Hugot JP, Durand P, Renaud F (2004) Evolutionary relationships between 15 Plasmodium species from new and old World primates (including humans): an 18S rDNA cladistic analysis. Parasitology 129:677–684

Lemey P, Salemi M, Vandamme A-M (eds) (2009) The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing. Cambridge University Press, Cambridge

Linder CR, Warnow T (2006) An overview of phylogeny reconstruction. In: Aluru S (ed) Handbook of computational molecular biology. Chapman & Hall/CRC Computer & Information Science, Boca Raton, FL

Linnaeus C (1758) Systema naturae. 10th edn., Stockholm

Martins EP, Hansen TF (1997) Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. Am Nat 149(4):646–667

Martins EP, Housworth EA (2002) Phylogeny shape and the phylogenetic comparative method. Syst Biol 51(6):873–880. doi:10.1080/10635150290155863

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. J Chem Phys 21:1087–1092

Minin V, Abdo Z, Joyce P, Sullivan J (2003) Performance-based selection of likelihood models for phylogeny estimation. Syst Biol 52 (5):674–683. doi:10.1080/10635150390235494

Moriyama EN, Powell JR (1997) Synonymous substitution rates in Drosophila: mitochondrial versus nuclear genes. J Mol Evol 45:378–391

Morlon H, Parsons TL, Plotkin JB (2011) Reconciling molecular phylogenies with the fossil record. Proc Natl Acad Sci 108(39):16327–16332. doi:10.1073/pnas.1102543108

Nakhleh L (2013) Computational approaches to species phylogeny inference and gene tree reconciliation. Trends Ecol Evol 28(12):719–728. doi:10.1016/j.tree.2013.09.004

Nei M, Kumar N (2000) Molecular evolution and phylogenetics. Oxford University Press, Oxford

Page RDM, Holmes EC (1998) Molecular evolution: a phylogenetic approach. Blackwell Publishing, Oxford

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

Pagel M, Meade A (2006) Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. Am Nat 167(6):808–825

Pagel M, Meade A, Barker D (2004a) Bayesian estimation of ancestral character states on phylogenies. Syst Biol 53(3):673–684. doi:10.1080/10635150490522232

Pagel M, Meade A, Barker D (2004b) Bayesian estimation of ancestral character states on phylogenies. Syst Biol 53(5):673–684

Paradis E (2011) Analysis of phylogenetics and evolution with R, 2nd edn. Springer, Berlin

Posada D, Buckley TR (2004) Model selection and model averaging in phylogenetics: advantages of akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst Biol 53(5):793–808. doi:10.1080/10635150490522304

R Development Core Team (2007) R: a language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. doi:http://www.R-project.orgS

Revell LJ, Reynolds RG (2012) A new Bayesian method for fitting evolutionary models to comparative data with intraspecific variation. Evolution 66(9):2697–2707. doi:10.1111/j.1558-5646.2012.01645.x

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574

Roquet C, Thuiller W, Lavergne S (2013) Building megaphylogenies for macroecology: taking up the challenge. Ecography 36:13–26. doi:10.1111/j.1600-0587.2012.07773.x

Rzhetsky A, Nei M (1992) A simple method for estimating and testing minimum-evolution trees. Mol Biol Evol 9:945–967

Saitou N, Nei M (1987) The neighbor-joining method—a new method for reconstructing phylogenetic trees. Mol Biol Evol 4(4):406–425

Sanderson MJ (1997) A nonparametric approach to estimating divergence times in the absence of rate constancy. Mol Biol Evol 14(12):1218–1231

Santos JC (2012) Fast molecular evolution associated with high active metabolic rates in poison frogs. Mol Biol Evol 29(8):2001–2018

Santos-Gally R, Gonzalez-Voyer A, Arroyo J (2013) Deconstructing heterostyly: the evolutionary role of incompatibility system, pollinators, and floral architecture. Evolution 67(7):2072–2082

Sibley CG, Ahlquist JE (1990) Phylogeny and classification of birds: a study in molecular evolution. Yale University Press, New Haven

Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22 (21):2688–2690. btl446 [pii] doi:10.1093/bioinformatics/btl446

Stone GN, Nee S, Felsenstein J (2011) Controlling for non-independence in comparative analysis of patterns across populations within species. Phil Trans R Soc B 366(1569):1410–1424. doi:10.1098/rstb.2010.0311

Symonds MRE (2002) The effects of topological inaccuracy in evolutionary trees on the phylogenetic comparative method of independent contrasts. Syst Biol 51:541–553

Thomas GH, Hartmann K, Jetz W, Joy JB, Mimoto A, Mooers AO (2013) PASTIS: an R package to facilitate phylogenetic assembly with soft taxonomic inferences. Methods Ecol Evol 4:1011–1017. doi:10.1111/2041-210X.12117

Wolfe KH, Sharp PM, Li W-H (1989) Rates of synonymous substitution in plant nuclear genes. J Mol Evol 29:208–211

Wu D, Jospin G, Eisen J (2013) Systematic identification of gene families for use as "markers" for phylogenetic and phylogeny-driven ecological studies of bacteria and archaea and their major subgroups. PLoS ONE 8(10):e77033. doi:10.1371/journal.pone.0077033

Zwickl DJ (2006) Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. University of Texas at Austin, Austin

Chapter 3
An Introduction to Supertree Construction (and Partitioned Phylogenetic Analyses) with a View Toward the Distinction Between Gene Trees and Species Trees

Olaf R. P. Bininda-Emonds

Abstract The dominant approach to the analysis of phylogenomic data is the concatenation of the individual gene data sets into a giant supermatrix that is analyzed en masse. Nevertheless, there remain compelling arguments for a partitioned approach in which individual partitions (usually genes) are instead analyzed separately and the resulting trees are combined to yield the final phylogeny. For instance, it has been argued that this supertree framework, which remains controversial, can better account for natural evolutionary processes like horizontal gene transfer and incomplete lineage sorting that can cause the gene trees, although accurate for the evolutionary history of the genes, to differ from the species tree. In this chapter, I review the different methods of supertree construction (broadly defined), including newer model-based methods based on a multispecies coalescent model. In so doing, I elaborate on some of their strengths and weaknesses relative to one another as well as provide a rough guide to performing a supertree analysis before addressing criticisms of the supertree approach in general. In the end, however, rather than dogmatically advocating supertree construction and partitioned analyses in general, I instead argue that a combined, "global congruence" approach in which data sets are analyzed under both a supermatrix (unpartitioned) and supertree (partitioned) framework represents the best strategy in our attempts to uncover the Tree of Life.

3.1 Introduction

A comparatively recent, but nevertheless fundamental insight within the field of comparative biology was the realization that it could only be done properly within a phylogenetic context (Felsenstein 1985b; see also Chap. 1), thereby disentangling

O. R. P. Bininda-Emonds (✉)
AG Systematics and Evolutionary Biology, IBU—Faculty V,
Carl von Ossietzky Universität Oldenburg, Carl von Ossietzky Strasse 9–11,
26111 Oldenburg, Germany
e-mail: olaf.bininda@uni-oldenburg.de

the similarity between species that arises via natural selection and convergent evolution versus that from their shared evolutionary history ("phylogenetic inertia"; sensu Harvey and Pagel 1991). Applying this form of phylogenetic correction thereby also acts as a statistical "fix" for any effects from other unmeasured variables. More recently, the use of well-resolved phylogenetic trees have helped to provide valuable insights into speciation and extinction rates (including their correlates and variation between and within groups), models of trait evolution, and community phylogenetics, among other fields. The key to performing all such analyses, naturally, is a reliable estimate of the phylogenetic history of the focal taxa, where great strides have been made within the last 25 years due in large part to the molecular revolution.

Despite many claims to the contrary, molecular phylogenetics has generally not uprooted our picture of the Tree of Life (Hillis 1987; Asher and Müller 2012) and many taxa have escaped the molecular revolution fairly unscathed (e.g., mammals or insects as a group). Furthermore, support for many phylogenetic hypotheses supposedly rooted on molecular data can also be found from morphological data. For instance, the molecular hypothesis that whales nest within even-toed ungulates rather than form the sister group to it was actually proposed at least as early as Beddard (1900) based on anatomical evidence (although admittedly largely ignored since then). Moreover, a recent study (Lee and Camens 2009) showed that many morphological data sets also contain substantial HIDDEN SUPPORT (see Box 3.1 for this and all other glossary entries as indicated in small caps) for otherwise conflicting molecular hypotheses of mammal phylogenetic relationships. Nevertheless, what the molecular revolution has unquestioningly provided is a plentiful, universal data source (i.e., DNA sequence data) that is becoming increasingly easy to tap into. Indeed, the advent and cost-effectiveness of next-generation sequencing means that DNA sequence data are often no longer limiting for phylogenetic purposes and are arguably becoming computationally, rather than financially prohibitive! A clear example here is the 1KITE project (http://www.1kite.org), with its goal of obtaining the entire transcriptomes of 1000 insect species covering all known orders, an amount of sequence data that would have been unthinkable a decade ago.

Box 3.1 List of abbreviations

HGT	Horizontal gene transfer
ILS	Incomplete lineage sorting
MLC	Multilocus coalescent model
MRP	Matrix representation with parsimony
OG	Outgroup
STK	Supertree Tool Kit

Accordingly, methodological discussions in molecular phylogenetics have long since shifted from issues of data quantity (e.g., if a limited number of taxa or characters is more detrimental with respect to accuracy; Graybeal 1998) to the best way to analyze the sequence data that are now so abundant. In this regard, the de facto standard is the total evidence or supermatrix approach, in which all the sequence data are concatenated into a single matrix and analyzed en masse. Proponents of this approach have championed it using both philosophical and methodological arguments. In the former case, the principle of "total evidence" is invoked in that the method uses all available data (sensu Kluge 1989). (Theoretically, nonsequence data that can also be accommodated in a matrix format (e.g., morphological characters) can also be included in the analysis; however, this is the exception rather than the rule.) In the latter case, simultaneous analysis facilitates the phenomenon of SIGNAL ENHANCEMENT (sensu de Queiroz et al. 1995)/HIDDEN SUPPORT (sensu Gatesy et al. 1999), whereby the concatenated data set might present a novel solution compared to the individual data partitions through the combination of the latter and effective upweighting of their consistent secondary signals. Importantly, analytical possibilities within a supermatrix framework have also kept pace, with analyses of over 50,000 taxa under a likelihood framework now being possible (e.g., Smith et al. 2011), including the possibility to apply separate models of evolution for each individual partition, even for disparate data types (e.g., DNA, amino-acid, and morphological data) (Stamatakis in press), thereby assuring a more optimal analysis of each data type or partition.

However, even within the possibilities offered by individual Bremer support analyses for each partition (partitioned Bremer support; Baker and DeSalle 1997) to visualize conflict among the different data partitions, the supermatrix approach tends to neglect that different genes often have different evolutionary histories and ones that can differ from that of the species. This fundamental gene tree/species tree conflict has been recognized since at least Maddison (1997) and derives from two main causes. The first problem is that individual genes essentially represent a statistical sample of the entire population (i.e., the genome) and so are subject to normal sampling artifacts. Thus, small genes might not possess a sufficient sample size in terms of the number of base pairs they contain to provide an accurate or well-resolved solution. Compounding this problem is that because DNA consists of only four nucleotides, it is subject to convergent evolution that typically confounds phylogenetic analyses as either SATURATION and/or LONG-BRANCH ATTRACTION (see Bergsten 2005) for fast-evolving genes along long branches. By contrast, extremely short branches, as are typically found in adaptive radiations, are also problematic because of the insufficient time to generate substitutions that provide evidence of the order of speciation events. Indeed, because such substitutions are more likely to derive from fast-evolving sites because of the short-time interval, this evidence is also more likely to disappear with time through subsequent substitutions at the same site and saturation.

Together, the above issues represent the normal, stochastic variation associated with any population estimate, possibly confounded by biases in the method of phylogeny reconstruction (e.g., long-branch attraction). A second, less appreciated
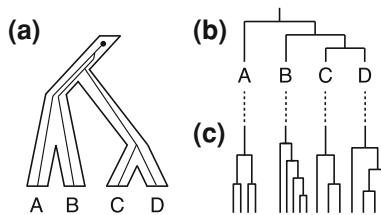
Fig. 3.1 The traditional representation of incomplete lineage sorting (ILS). **a** The species tree is represented by the outline and contains a gene tree (*thin lines*). In this case, the gene tree conflicts with the species tree and gives a wrong estimate of it (**b**). This problem will not disappear with time given that the terminal taxa of (**b**) comprise the common ancestors of those in (**c**). ILS is more prevalent during rapid speciation in large populations, when the time to coalescence for the gene tree is less than that for speciation

problem is that the evolutionary history of a gene can truly differ from that of the species due to any number of natural processes such as horizontal gene transfer (HGT), recombination or incomplete lineage sorting (ILS, also known as deep coalescence; see Fig. 3.1). Although these phenomena were believed originally to be relatively rare and/or confined to otherwise difficult taxa like microbes, there is a growing realization that both HGT and ILS might indeed be both more common and widespread than we had thought previously. Indeed, given the right set of evolutionary conditions (e.g., rapidity of speciation events), the number of gene trees that conflict with the species tree can outnumber the number that agree with it, even for species trees containing as few as five species (Degnan and Rosenberg 2006). Both phenomena are also particularly insidious because they imply that our gene trees are accurate even though they misrepresent the species tree! Again, recent, rapid speciation events are particularly problematic, especially when they occur sequentially (Rosenberg 2013) and in large populations with low rates of genetic drift because the speciation rate exceeds the coalescent rates of the different genes in this so-called anomaly zone (Degnan and Rosenberg 2006), thereby facilitating ILS (Steel and Rodrigo 2008; Edwards 2009). Even more worrisome is that the misleading effects of both HGT and ILS do not necessarily disappear with time. In the case of ILS, because the daughter species arising from the speciation event represent the common ancestors of future higher-level clades (e.g., the "orders" within mammals), a misleading species tree from the past can translate into a misleading ordinal tree in the future (see Fig. 3.1c).

As something of an aside, the same artifacts can potentially arise in the absence of ILS through the process of speciation itself, which can create paraphyletic daughter species. A classic example is the origin of the polar bear (*Ursus maritimus*). Although recent studies based on nDNA markers now indicate it to be an ancient lineage forming the sister group to brown bears (*Ursus arctos*) (Hailer et al. 2012; Miller et al. 2012), it was believed until very recently that this species, based on studies of mtDNA, arose from isolated populations of brown bear from the Admiralty and Baranof Islands of the Alexander Archipelago of southeastern

Alaska about 150,000 years ago (e.g., Lindqvist et al. 2010). Were the latter scenario indeed true (which is undoubtedly generally the case for many other species), then some individuals/populations of brown bear are more closely related to polar bears than to other members of their own species, which, depending on the pattern of future speciation and extinction in the brown-bear lineage, could give rise to ILD-like knock-on effects.

As such, there have been recent attempts to move away from a pure supermatrix approach to ones that can potentially better accommodate such instances of "gene-tree heterogeneity" (sensu Edwards 2009) by focusing on the gene tree as the fundamental unit of analysis rather than individual nucleotides or amino acids. In so doing, it is recognized that although gene trees and species trees are closely related evolutionarily, they nevertheless derive from distinct evolutionary processes (Liu et al. 2010). In essence, these arguments are merely the latest thoughts in a long-standing debate as to whether it is more desirable to automatically combine data or to perform some form of partitioned analysis (see Chippindale and Wiens 1994).

Against this backdrop, the goal of this chapter is to outline and describe two such frameworks for partitioned analyses: the now "traditional" supertree approach and the more recent multilocus coalescent (MLC) model that explicitly builds on coalescent and population genetic theory to derive a species tree from a set of potentially conflicting gene trees. Both approaches are united in having their analytical focus at the level of a set of input trees and, although this was never advanced originally as a justification for traditional supertrees, thereby possess the potential to account for any gene-tree heterogeneity. More controversially, both for expediency and because of the unquestionably strong parallels between the two frameworks, I will refer to them collectively as "supertrees".

This chapter is structured as follows. First, I initially provide a short historical perspective of the supertree framework before providing a summary of both traditional supertree methods and the newer methods based on the MLC model (both summarized in Table 3.1). In so doing, I hope to show the similarities between these two "classes" of methods as well as to point out that MLC-based methods are not the only supertree methods to include an explicit evolutionary model. Second, I briefly address previous criticisms of the supertree framework, especially in relation to the supermatrix framework. However, here and throughout, apart from general comparisons to the supertree framework, I will largely refrain from discussing details of the mechanics of a supermatrix analysis given the overwhelming prevalence of (and therefore likely familiarity with) this technique. An excellent summary of general phylogenetic tree building, which forms the backbone of the supermatrix framework (as well as the derivation of individual gene trees), is also provided in Chap. 2 of the volume. Finally, I provide a rough guide as to how to perform a supertree/partitioned data analysis. Given the vast array of supertree methods available, this guide is purposely agnostic in the sense that it does not advocate any one method, but concentrates instead on the various issues that must be considered at each step in the process.

Table 3.1 A summary of the supertree methods listed in the main text, including some brief notes on their properties (where known) and implementations

Method	Notes	Implementation
Average consensus	A method that can explicitly account for branch-length information among the set of source trees when constructing the supertree.	CLANN (http://bioinf.nuim.ie/clann/) (Creevey and McInerney 2005)
Gene-tree parsimony	Derives the supertree based on the method of reconciled trees and therefore explicitly accounts for biological processes like gene duplication and loss.	DupTree (http://genome.cs.iastate.edu/CBL/DupTree/) (Wehe et al. 2008); iGTP (http://genome.cs.iastate.edu/CBL/iGTP/) (Chaudhary et al. 2010)
Matrix representation	A class of methods whereby the topology of a source tree is encoded as a matrix using additive binary coding. The resulting matrix derived from all source trees can be analyzed using virtually any optimization criterion (e.g., MP, NJ, ML, BI, or compatibility).	Among others: CLANN; Rainbow (http://genome.cs.iastate.edu/CBL/download/) (Chen et al. 2004); SuperMRP.pl (http://www.uni-oldenburg.de/ibu/systematik-evolutionsbiologie/programme/) (Bininda-Emonds et al. 2005); SuperTree (http://www2.unil.ch/phylo/bioinformatics/supertree) (Salamin et al. 2002); Supertree Tool Kit (http://sourceforge.net/projects/stck) (Davis and Hill 2010)
MRF (matrix representation with flipping)	A matrix representation method with a specific implementation. The optimization criterion of "flipping" involves finding the fewest number of 0 → 1 or 1 → 0 changes in the matrix required to produce a matrix without conflict.	HeuristicMRF2 (http://genome.cs.iastate.edu/CBL/download/); Rainbow
MINCUTSUPERTREE	Fast, polynomial time method that preserves nestings among the source trees and where the supertree displays each source tree. Because it preserves nestings and not clades, it is akin to Adams consensus and the supertree cannot necessarily be interpreted phylogenetically.	Supertree (http://darwin.zoology.gla.ac.uk/%7Erpage/supertree) (Page 2002)
Modified MINCUTSUPERTREE	Modification to the previous method to achieve greater resolution.	Supertree; Rainbow

Table 3.1 (continued)

Method	Notes	Implementation
MULTILEVELSUPERTREE	Accounts for both horizontal and vertical overlap among the set of source trees and so can account for nested taxa among the set of source trees.	MLS (http://www.atgc-montpellier.fr/supertree/mls/) (Berry et al. 2013)
Multilocus coalescent models	A class of methods that rely on coalescent theory to obtain the supertree from the set of gene trees and so explicitly account for biological processes such as ILS. Some methods can account for branch-length information within the gene trees.	MP-EST (https://code.google.com/p/mp-est/) (Liu et al. 2010) Phybase (https://code.google.com/p/phybase/) (Liu and Yu 2010)
PhySIC and PhySIC_IST	Derive a supertree showing relationships that do not contradict any of those on the source trees and are induced by them (in both cases, singly or jointly).	Webservers available at: http://www.atgc-montpellier.fr/physic/ http://www.atgc-montpellier.fr/physic_ist/ (Ranwez et al. 2007; Scornavacca et al. 2008)
Quartet puzzling	Examines all possible quartets to determine the most likely one for any set of four taxa. The latter are then essentially combined into the final tree. Used for the analysis of DNA data, but the principle could be applied in a supertree framework.	TREE-PUZZLE (http://www.tree-puzzle.de) (Schmidt et al. 2002)
Quartet supertrees	Breaks source trees down into all their possible quartets and builds a supertree from the latter based on their frequencies across the set of source trees.	Quartet suite (http://genome.cs.iastate.edu/CBL/download/) (Piaggio-Talice et al. 2004)
SuperFine	A meta-method that can theoretically be used to speed up ("boost") any existing supertree method.	SuperFine (http://www.cs.utexas.edu/~phylo/software/superfine/) (Swenson et al. 2012)

All would be used in Step 3 of the guide presented in Sect. 3.4.3 of the text.
Note that many of the implementations have not been recently updated and so might not run on the latest operating systems

3.2 The Supertree Framework

Supertrees are essentially as old as systematics itself, where our vision of the Tree of Life as a whole was essentially patched together from many smaller subtrees, often using a form of taxonomic substitution. In this, the terminal taxa in a higher-level tree were simply substituted for the nested tree showing the relationships within that taxon. Thus, for a tree of the vertebrate classes, the taxon Mammalia could be replaced by an ordinal-level tree of this group, for example, and so on. Although this technique is still in use today to provide us with some picture of the Tree of Life as a whole, it is distinctly limited in that it requires us to choose some "best" tree at each level and so make a subjective judgment among the many, possibly conflicting options.

A more objective foundation for supertrees essentially dates to 1986, when the mathematician Allan Gordon proposed a generalization of the well-known strict consensus method that could be applied to a set of trees that differed in the terminal taxa they contained (Gordon 1986). For various largely methodological reasons, the solution was largely unworkable and/or ignored (see Bininda-Emonds 2004b), and it was only in 1992 that the next breakthrough was achieved by Baum (1992) and Ragan (1992), who independently described the method now known as matrix representation with parsimony (MRP). Building on the one-to-one correspondence between a tree and its binary equivalent in matrix form ("MATRIX REPRESENTATION"; Ponstein 1966) (see Fig. 3.2), Baum and Ragan each hit upon the idea of concatenating the individual matrix representations of a set of source trees
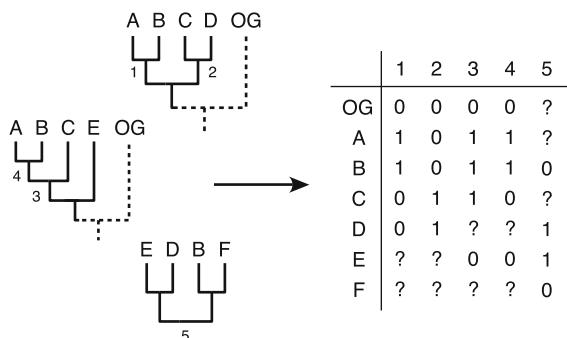
Fig. 3.2 Matrix representation of a set of three gene trees. Using additive binary coding, the nodes of any given gene tree can be represented in matrix format in turn. For a focal node, terminal taxa that are descended from that node receive *1*, taxa that are not but are present on the gene tree receive *0*, and all other taxa receive? (e.g., character *2* which represents node *2*). To root the analysis, a fictitious outgroup (OG) comprising all *0*s is added to the base of each gene tree. If a distinction is made between rooted and unrooted gene trees, the OG can also receives? for unrooted source trees (character *5*; see Bininda-Emonds et al. 2005). For any single tree, there is a one-to-one correspondence between it and its matrix representation. To derive the supertree that best fits to the set of gene trees, the entire matrix is then analyzed using any desired optimization criterion (typically parsimony) after which the OG is subsequently removed

into a single (super)matrix and then analyzing the latter with parsimony to derive a "supertree". The potential of the method was quickly realized by Purvis (1995a), who combined numerous estimates of primate phylogeny taken from the literature to derive the first, complete species-level evolutionary tree for the group based on an objective, robust methodology. From there, the field exploded, both in terms of the supertrees themselves that were being generated as well as the supertree methods used to obtain them. A now highly outdated list on both counts can be found in Bininda-Emonds (2004a) and many new trees and methods have been developed since then. It nevertheless remains that MRP is by far and away the most popular of the supertree methods.

3.2.1 Traditional Supertrees

With the growing number of methods (see Table 3.1), supertrees are becoming increasingly difficult to summarize meaningfully as a group as well as to categorize with respect to their methodologies. The supertree framework has historically been seen as a generalization of that for consensus trees, which requires identical taxon sets among the source trees. However, the analogy only holds so far in that most supertree methods do not have clear consensus equivalents (including MRP) and many popular consensus methods did not have a corresponding supertree one until comparatively recently (e.g., majority-rule supertrees; Cotton and Wilkinson 2007). In the end, perhaps, a supertree is now best summarized as the summary tree derived from a set of source trees that need not have identical taxon sets. Under this definition, supertrees remain a generalization of consensus trees, but can extend beyond this as well. An alternative, but not mutually exclusive, interpretation is that the summary tree obtained from a supertree analysis represents the "best fit" to the set of source trees according to some objective function (Thorley and Wilkinson 2003; Bruen and Bryant 2008). In most cases (including MRP), this objective function is unknown, but some supertree methods have been developed explicitly with an objective function in mind, including majority-rule (minimizes partition metric; Cotton and Wilkinson 2007) and maximum-likelihood supertrees (minimizes error function among source trees; Steel and Rodrigo 2008) as well as MINCUTSUPERTREE (minimizes sum of triplet distances; Wilkinson et al. 2004).

A clear subcategory of supertrees are those that, like MRP, rely on an explicit intermediate step of building a matrix of pseudo-characters, with each pseudo-character representing a node on a particular source tree. In a sense, the combined matrix functions as a table of the bipartition frequencies among the set of source trees, where a bipartition splits an unrooted tree into two taxon sets (e.g., for the bipartition AB|CD, removing a branch on the tree will result in two subtrees, one with taxa A and B and the other with taxa C and D). The matrix can then be analyzed using any preferred optimization criterion. Although parsimony remains by far the method of choice here (as in MRP), other suggested methods include compatibility (Ross and Rodrigo 2004), flipping (Chen et al. 2003), Bayesian

inference of the bipartition frequencies (Ronquist et al. 2004), or, most recently, maximum likelihood with a two-state/parsimony character model (Nguyen et al. 2012). A variant on these methods is the average consensus method (Lapointe and Cucumel 1997), whereby the matrix to be analyzed consists of the sum of the path-length distances between all pairs of taxa among a set of gene trees, with some estimate of these distances for pairs of taxa that do not co-occur on any single tree (Lapointe and Levasseur 2004).

A long-standing critique of traditional supertree methods is that, although arguably reasonably accurate empirically and in simulation, most are not based on any explicit model of biological evolution (Liu et al. 2010) and so are better classed as nonparametric methods (Liu et al. 2009a) and/or the properties of most remain uncharacterized (see Wilkinson et al. 2004). Indeed, MRP represents the poster child that has attracted the most attention in this regard. Despite its long-standing popularity and use, the objective function of MRP remains unknown and the little that is known about its properties is worrisome. For example, it was known almost from the outset that MRP, like many other supertree methods, gives more weight to larger source trees (i.e., is not "sizeless"; Purvis 1995b; Wilkinson et al. 2004) (although it actually favors larger subtrees rather than trees as a whole; see Bininda-Emonds and Bryant 1998); other potentially undesirable properties are summarized in Wilkinson et al. (2004). Nevertheless, the fact that MRP shows reasonable accuracy in practice and can even outperform equivalent supermatrix analyses in simulation (Bininda-Emonds and Sanderson 2001) show that its deficiencies are either not that severe and/or only arise in extreme cases.

That being said, a few supertree methods have been designed explicitly to fulfill some properties identified by Steel et al. (2000) and Wilkinson et al. (2004) as being desirable when combining trees ("desiderata"; sensu Wilkinson et al. 2004). In some ways, this can be viewed as part of the objective function of these methods. For instance, MINCUTSUPERTREE (Semple and Steel 2000) and its derivative modified MINCUTSUPERTREE (Page 2002) output supertrees that can be found in POLYNOMIAL TIME, preserve nestings and binary trees found among all source trees, display all input trees if the latter are compatible, and are independent of the input order of the trees (Semple and Steel 2000). However, by preserving nestings, rather than clades, among the sets of source trees (Semple and Steel 2000), the MINCUTSUPERTREE tends to resemble the Adams consensus (Adams 1972, 1986) of the source trees, meaning that the result cannot always be interpreted phylogenetically. For instance, MINCUTSUPERTREE will only preserve the nesting information that A and B are a part of a larger cluster ABCD, without any statement as to the relationship between A and B themselves. Thus, even if A and B form sister taxa in the resulting supertree it cannot automatically be assumed that they do indeed form a clade (because MINCUTSUPERTREE does not preserve clades).

Other examples of supertree methods designed to meet certain properties a priori are PhySIC (Ranwez et al. 2007) and PhySIC_IST (Scornavacca et al. 2008), which ensure that the resulting supertree displays all relationships that are induced by and are not contradicted by the set of source trees, either alone or in combination. The latter method builds on the former by removing highly conflicting

source trees in the hopes of obtaining a better resolved supertree. Together, both methods are perhaps a direct answer to methods like MRP, which theoretically can output relationships that are contradicted by every source tree (see Bininda-Emonds and Bryant 1998) although this appears to be extremely rare in practice (Bininda-Emonds 2003).

3.2.2 Multilocus Coalescent "Supertrees"

A more recent advance toward supertree methods based on evolutionarily sound models—and on the potential distinction between gene trees and the species tree in particular—is the MLC model, which builds theoretically on Rannala and Yang's (2003) characterization of the likelihood function of the species tree under a multispecies coalescent via two probability distributions (Liu et al. 2009a). The first, $f(\mathbf{D}|\mathbf{G})$, describes the probability of deriving a particular gene tree ($\mathbf{G}$) given a set of sequence data ($\mathbf{D}$) and represents the same likelihood function used routinely in molecular phylogenetics. The second, $f(\mathbf{G}|S)$, describes the probability of observing a gene tree given a particular species tree (S) and derives from the multispecies coalescent. Essentially, for a species tree with well-defined clades separated by long branches (i.e., divergence times), the majority of gene trees will resemble the species tree and gene-tree heterogeneity will be low. However, when the species tree contains one or more regions with short branch lengths, the probability for gene-tree heterogeneity in these anomaly zones increases and many more, different gene trees are expected.

Practical implementations of the MLC model, however, are more indirectly related to these probability distributions. Indeed, at least one procedure has been termed as a "maximum pseudo-likelihood approach" by its authors (Liu et al. 2010). Given a set of gene trees (essentially component $f(\mathbf{D}|\mathbf{G})$ from above), one form of the MLC model derives a distance matrix between all pairs of terminal taxa based on their coalescent events across the set of gene trees. For any given cell, the distance value is given either by (1) the minimum number of ranks (nodes) across the set of gene trees until the taxa share a common ancestor/coalesce (GLASS distance; Mossel and Roch 2007), (2) the average number of ranks until they do so (STAR distance; Liu et al. 2009b) or (3) the average coalescence time (STEAC distance; Liu et al. 2009b). Thus, whereas the first two distances only account for topological information within the gene trees (and are therefore only "partially parametric"; Song et al. 2012: 14943), the last can incorporate branch-length information directly when it is present. Finally, the distance matrix is analyzed via a distance method like NJ to derive the species tree (component $f(\mathbf{G}|S)$ from above). By contrast, a second implementation of the MLC model, MP-EST (Liu et al. 2010) derives the frequencies of all triplets of taxa from the set of gene trees (together with path-length information) to obtain the topology and branch lengths of the species tree in a pseudo-likelihood framework, again representing a "partially parametric" method. Both sets of methods appear to perform

well under conditions of gene-tree heterogeneity where equivalent supermatrix analyses become statistically inconsistent (Wu et al. 2013).

Although it has not been recognized to date, the two implementations of the MLC model have clear connections to existing traditional supertree methods. For example, the distance-based MLC methods bear strong resemblances with the average consensus supertree method in that both explicitly incorporate branch-length information from the gene trees, even if only indirectly in the form of ranks. Likewise, MP-EST shows similarities with quartet puzzling (Strimmer and von Haeseler 1996) or quartet supertrees (Piaggio-Talice et al. 2004), albeit with MP-EST requiring rooted gene trees (and hence using triplets) instead of the unrooted framework (and thus quartets) employed by the latter two methods. (Quartet puzzling also proceeds directly from the DNA sequence data without explicit regard to gene trees. However, it could be modified from this supermatrix format to work in a gene-tree context.) More generally, the explicit use of an underlying biological model also characterizes the gene-tree parsimony method (Cotton and Page 2004), which has been recognized as a supertree method and uses reconciled trees (Goodman et al. 1979; Page 1994) to account for possible discrepancies between the gene trees and the species tree as a result of processes including HGT and gene duplication and loss.

Nevertheless, by being explicitly couched within a coalescent framework and building on the likelihood function of Rannala and Yang (2003), the MLC methods differ from most other supertree methods in being based on explicit biological models and phenomena. For instance, the MLC model assumes, among other things (see Liu et al. 2009a), constant population sizes through time, random mating, no gene flow, and no HGT, and thus the predominance of ILS as the cause of gene-tree heterogeneity. Although many of these assumptions are unrealistic, the MLC methods are apparently robust to minor rates of HGT and could, in theory, be easily expanded to account for both this and gene flow (Liu et al. 2009a).

MLC-based methods also possess a distinct advantage in that they are very fast compared to most other supertree methods (except for polynomial-time methods like MinCutSupertree) once the input trees have been calculated. As shown by Liu et al. (2010), runtimes are on the order of seconds for problem sizes of 80 gene trees each comprising 20 taxa, both for STAR-based analyses as well as those using ML-EST, albeit with the latter being demonstrably slower. The speed accrues either from the use of NJ as an optimization criterion or the pseudo-likelihood framework compared to the NP-complete algorithms (e.g., parsimony or likelihood) typically used by traditional supertree methods. However, even in the latter case, tremendous speed gains have been achieved by implementing supertrees in a divide-and-conquer framework, in which the supertrees represent more of a search strategy than the end product of the analysis (see Bininda-Emonds 2010). Here, the general idea is to take a large, computational demanding problem (e.g., a large multigene data set of thousands of taxa) and to break it down into many smaller, overlapping data sets that are more tractable because of their small size. The resulting trees from the latter data sets are then combined as a supertree, which can then be further resolved on the basis of the entire data set (Roshan et al. 2004). This general strategy, which also underlies

quartet puzzling, has most recently been implemented in SuperFine (Swenson et al. 2012), a so-called meta-method designed to boost the speed of existing supertree methods like MRP. Indeed, the method does appear to deliver more optimal supertrees in a reduced amount of time compared to nonboosted analyses (Swenson et al. 2012; Nguyen et al. 2012), but still at best only on a par in terms of speed and accuracy with equivalent supermatrix analyses (Swenson et al. 2012; Nguyen et al. 2012). In this, the problem with the divide-and-conquer approach appears to lie with the final resolving step, which is based on the full data set and is therefore subject to the same size-based tractability problems (Bininda-Emonds 2010).

3.2.3 Accounting for Vertical Taxonomic Overlap

A feature shared by all the above methods is that they essentially only account for horizontal overlap among the gene trees (i.e., among the terminal taxa). As such, the terminal taxa must all occur at the same taxonomic level (e.g., species in the case of gene trees) or minimally cannot be nested within one another. Thus, the case where a source tree possessed the terminal taxon Mammalia and another possessed *Homo sapiens* would result in a supertree where these two taxa would, at best, be sister groups, despite the latter clearly nesting within the former. Recalling to some degree the process of taxonomic substitution characterizing informal supertree methods, MULTILEVELSUPERTREE (Berry et al. 2013) is able to simultaneously account for both horizontal and vertical overlap among the source trees, the latter representing the nested, higher-level relationships implicit among the set of source trees. Moreover, the program is also able to infer the latter from information among the source trees themselves, such that it is not necessary to provide a reference taxonomy providing the nested sets of relationships. Although MULTILEVELSUPERTREE would appear to be of use when combining source trees out of the literature, this traditional use of supertrees is rapidly falling by the wayside and it is not clear if its ability to also accommodate vertical overlap will provide any benefit for gene trees based on DNA sequence data, which normally all have species as terminal taxa.

3.3   Criticisms of Supertrees

Even when couched within the context of explicitly accommodating gene-tree heterogeneity, the supertree framework has been highly criticized and remains controversial (e.g., see the exchange between Gatesy and Springer (2013) and Wu et al. (2013) for MLC-based methods). The primary areas of criticism include (1) the potential for duplication of data between the source partitions, (2) the black-box nature of most supertree methods and MRP in particular, and (3) the fact that the methods are a form of meta-analysis and thus one step removed from the primary character data.

Data duplication does indeed represent a potential problem area within a supertree framework as was elegantly shown by Gatesy et al. (2002) for the supertree analysis of mammalian families by Liu et al. (2001). For instance, the same genes (if not the same sequences) are often used for separate phylogenetic analyses, often in combination with other genes. A cogent example here is cytochrome b, which represents by far the most widely sampled gene for mammals to date and one that is often used for phylogenetic analyses within the group. As such, it often comprises part of the data set underlying different phylogenetic trees for mammals, meaning that these trees are nonindependent of one another. Thus, constructing a supertree for mammals by simply collecting and combining all published trees for the group means that cytochrome b would have an unduly greater influence on the end result compared to other genes and sources of character data.

Indeed, many early supertree studies ran afoul of this problem before it was so forcefully pointed out by Gatesy et al. (2002). Fortunately, data duplication is a largely historical problem that can be mitigated today by more careful selection of the source trees and/or by complicated weighting schemes designed to address it (e.g., Nyakatura and Bininda-Emonds 2012). More generally, this criticism is largely obsolete when supertrees are used in an explicit gene-tree framework, where each gene tree is present only once within the data set. Even so, it should be remembered that even the subdivision of the genome into individual genes is to some extent subjective, with our concept of "genes" having become increasingly blurred with increased knowledge of the tremendous degree of complexity underlying the genome (e.g., via recombination, exon shuffling, HGT, and alternative splicing, among other processes). Instead, of note here are newly developed methods like PARTITIONFINDER (Lanfear et al. 2012), which use data-driven, information-theoretic metrics to more objectively reveal partitions within a data set (within the bounds of a set of a priori user-defined partitions). However, it remains to be seen how well these partitions match up with those expected under a gene-tree heterogeneity scenario largely driven by ILS (i.e., classic gene trees). The finding that individual genes are composed of several partitions (e.g., according to codon position in protein-coding genes or stems vs. loops in rDNA genes) would not be problematic, but instead serve to improve our estimate of the individual gene trees. By contrast, the sharing of partitions across genes might force us to rethink our notion of gene trees entirely.

The remaining two criticisms of supertrees are to some degree linked and mirror that of Liu et al. (2010) in claiming that traditional supertree methods do not resolve conflict among the source trees with respect to explicit evolutionary events (Gatesy and Springer 2004). However, this is no longer the case and several supertree methods, such as gene-tree parsimony and the MLC-based methods, now exist that fulfill this criterion. It is important to remember, however, that the supermatrix and supertree methods do operate at different hierarchical levels (DNA sequence data vs. gene trees, respectively; Bininda-Emonds 2004c) such that each will be accommodating different sets of evolutionary events (e.g., character-state transformations vs. HGT or ILS, respectively). Moreover, through

their focus at the level of the gene tree, only methods like gene-tree parsimony and the MLC-based methods have the potential to account for processes like ILS, which, when frequent enough, have been demonstrated in simulation to impact on the accuracy of supermatrix methods to the point of them being statistically inconsistent (Wu et al. 2013).

3.4 A Primer to Supertree Construction

The following represents a rough guide to the process of creating a supertree and is also illustrated in Fig. 3.3. It takes its form both from my own experiences and from their formalization and extension in the excellent Supertree Tool Kit (STK) of Davis and Hill (2010). Given the huge variety of supertree methods and choices available, the guide is not intended to be either exhaustive or dogmatic. Other, often unnamed, variations on this framework are conceivable and should be explored and not excluded a priori. A simple, worked example to be used as a jumping-off point can be found in the OPM.

3.4.1 Step 1: Obtaining the Source Trees

Much of the previous discussion has centered on the concept of gene trees, with the implication that they have been obtained directly via phylogenetic analysis of primary molecular sequence data by the researcher. These data can derive either from de novo sequences generated by the researcher and/or from online resources such as GenBank. Indeed, in the latter case, numerous phylogenetic pipelines now exist (see Bininda-Emonds 2011) for the express purpose of mining GenBank and other similar resources for homologous sequence data.

However, gene trees represent only one source of data potentially available under a supertree framework. Because the raw data of a supertree analysis is a phylogenetic tree, any statement of phylogenetic relationship that can be expressed as a bifurcating tree can be included in the analysis. It was this very principle that underlay the earliest empirical supertree studies in which source trees were mined from the literature and either encoded directly in matrix format or as nexus-formatted tree statements for later processing. The online archiving of phylogenetic trees through resources like TreeBASE (www.treebase.org; Sanderson et al. 1994) merely represents the modern and more convenient extension of the traditional paper-based sifting of the literature. Although the inclusion of literature data is quickly falling out of favor in the era of molecular phylogenetics, it remains that it provides access to more of the global phylogenetic database and data that would be otherwise difficult to include in a supermatrix framework. The latter includes not only older molecular data such as DNA–DNA hybridization or isozymes,
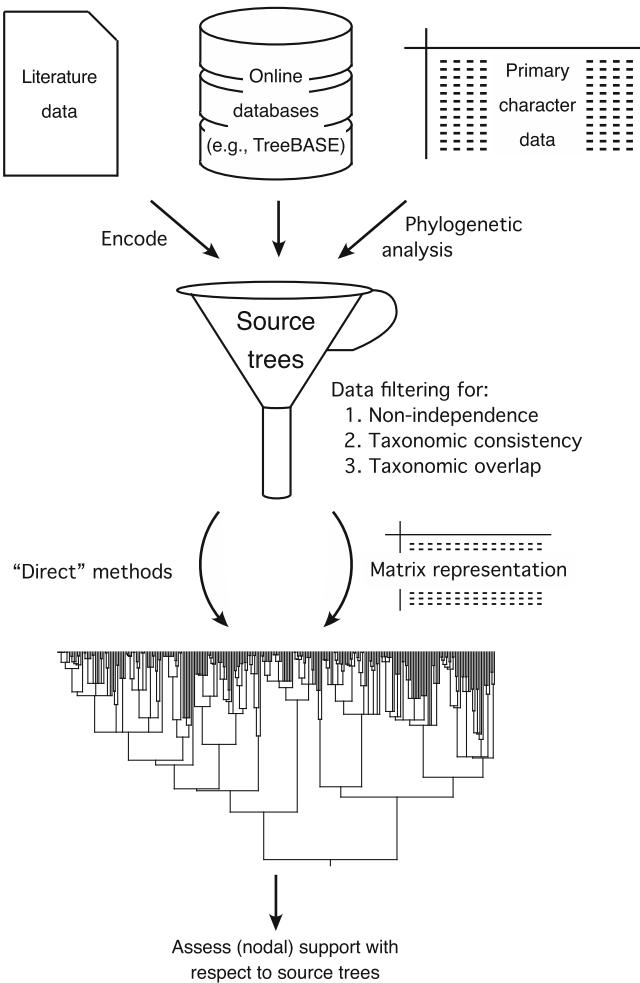
Fig. 3.3 Flow diagram illustrating the general framework of a supertree analysis. Particularly crucial is the middle, filtering step, which acts as a measure of quality control for the source trees derived from any or all of the literature, online databases, or primary character data. Thereafter, any supertree method of choice can be applied to the filtered set of source trees. Adapted from Bininda-Emonds et al. (2004) and Davis and Hill (2010)

but also morphology and evidence from rare genomic changes (Rokas and Holland 2000), the signals of both of which threaten to be swamped by the much more numerous molecular sequence data. Thus, in some respects, a supertree framework can better accommodate the principle of total evidence (i.e., using as much data as possible) than can a supermatrix framework (Bininda-Emonds et al. 2003). That being said, the inclusion of literature data does harbor particular difficulties that are addressed in the next step of the process.

3.4.2 Step 2: Filtering the Source Trees for Data-Quality Assurance

This crucial step was inspired initially by the paper of Gatesy et al. (2002), who elegantly documented several weaknesses with respect to data quality in several previously published empirical supertree studies (see above). Although Gatesy et al. (2002) took extended aim at the supertree framework in general, it remains that their paper is essentially about quality assurance in phylogenetic analysis in general and not for any method or framework in particular.

That being said, a supertree framework, especially one that incorporates literature data can present special problems in this regard, again because of the disconnect between the primary character data and the source trees that provide the raw data for the supertree analysis. Because the latter are often mute with respect to the former, a greater potential for the duplication at the level of the primary character data exists within a supertree framework (see above). However, as pointed out above, careful diligence, perhaps in combination with complicated weighting schemes to account for any data duplication, will often be sufficient to ameliorate this potential problem. Within a pure gene-tree framework, this problem is unlikely to occur or, at least, will have the same impact as on the equivalent supermatrix studies that could be performed on the fundamental data set. The issue of data quality, and which source trees to actually include in the analysis, is more thorny with arguably no correct answer. Whereas some investigators will be comfortable including taxonomies as source trees (ignoring the fact that it might be based in part on data also used in other source trees and so represent a case of data duplication), others will reject this possibility categorically. As with any scientific study, it is important in such cases to be open and to make the data available for other researchers to replicate the study under their preferred set of conditions. A sensitivity analysis can also be envisaged, whereby source trees of arguably lower quality are either downweighted or removed from the analysis to ascertain what their impact on the supertree topology is.

In a subsequent step, it is important to ensure consistency among the taxonomic labels among the set of source trees, especially for those methods that only account for horizontal overlap among the terminal taxa. Although MULTILEVELSUPERTREE, by also accounting for vertical overlap among the source trees, can avoid problems of nested taxa, a check for taxonomic consistency is necessary here as well to ensure that the same taxa are not present as different synonyms (e.g., Mammalia vs. just "mammals") in different source trees. The issue of taxonomic consistency was first raised by Bininda-Emonds et al. (2004), who also present different, general solutions to the overall problem, which can be implemented either through synonoTree.pl (Bininda-Emonds et al. 2004) or the STK (Davis and Hill 2010). Finally, although this general problem will be more rare in a pure gene-tree context, it can also be relevant here (e.g., GenBank often indexes sequence information separately for a species and its subspecies).

A final check, and one that is neatly implemented in the STK, is to assess the degree of taxonomic overlap among the set of source trees. Minimally, a supertree analysis requires that each source tree overlaps with at least one other by two terminal taxa. (This requirement is loosened through the vertical overlap recognized by MULTILEVELSUPERTREE.) Where such overlap does not occur, the resulting supertree should be completely unresolved (within the bounds of the optimization criterion used) because the taxa from one nonoverlapping group can cluster equally optimally with those from another group. This problem is easily ameliorated by either removing nonoverlapping source trees (or running separate supertree analyses for nonoverlapping sets of trees) or by including a "seed tree" (sensu Bininda-Emonds and Sanderson 2001) that contains most if not all the taxa among the set of source trees and so provides a scaffold for the analysis. The seed tree is often derived from a taxonomy, with the poor resolution of such entities meaning that the seed tree provides minimal clustering information of its own. The impact of the seed tree, the use of which is controversial, can be minimized further by down-weighting it within the analysis compared to the other primary source trees.

3.4.3 Step 3: Obtaining the Supertree

This represents the most obvious and direct of the four steps in performing a supertree analysis. However, as the previous Sect. 3.2 makes clear, the sheer and still growing variety of supertree methods (see also Table 3.1) can make the selection of the final method difficult. MRP remains by far the method of choice; however, this seems to obtain more for historical considerations rather than the method being demonstrably superior to any alternatives. Therefore, perhaps merely for reasons of comparability with other supertree studies, an MRP analysis is to be recommended. Nevertheless, other methods should also be explored, either because of their arguably better accuracy and/or because of their more desirable properties or objective functions.

It is in this third step where weighting is employed, not only to account for potential data duplication, but also for potential differences in the robustness/quality of the different source trees. (Early attempts to employ weighting to counteract the apparent size bias of MRP (Ronquist 1996) were ultimately unsuccessful because MRP does not favor larger source trees per se, but overlapping parts of those source trees (Bininda-Emonds and Bryant 1998), making any weighting scheme impossible to implement with large numbers of source trees with different degrees of overlap among them.) A simple solution here is to simply replicate entire source trees proportional to some measure of their inferred quality.

When weighting for source-tree "quality", however, it is important to recognize that phylogenetic relationships within any given tree can also differ in support, with some clades being comparatively better supported than others. In a default supertree analysis, where only the topology of the source trees is used, this information is completely lost, an early criticism of the supertree framework as a

meta-analysis (e.g., Gatesy and Springer 2004; but see above). A simple solution in this regard for matrix representation-based methods at least is to weight each pseudo-character by the inferred support for the node in the source tree that it encodes [e.g., according to its nonparametric bootstrap frequency (Felsenstein 1985a) or Bremer support (Bremer 1988)]. Indeed, although this form of weighting still cannot account for hidden support among the primary data partitions, simulation studies have shown that doing so improves the accuracy of MRP supertrees, often to the point where the supertree analysis slightly outperforms an equivalent supermatrix analysis (Bininda-Emonds and Sanderson 2001). Important here, however, is to ensure that the weighting schemes are comparable among the source trees (e.g., not a combination of bootstrap frequencies and Bremer support values); however, this should not be a problem for gene trees generated de novo from public databases like GenBank. For other supertree methods where there is no direct connection with the individual clades on a given tree, some form of clade duplication proportional to inferred support can also be envisaged.

Finally, it is important to realize that because all supertree methods ignore the data underlying the source trees, this third step essentially delivers a tree topology only. With the possible exception of the average consensus method, any branch lengths on the supertree are either essentially meaningless (e.g., MINCUTSUPERTREE or gene-tree parsimony) or cannot be interpreted phylogenetically (e.g., matrix representation methods). This is especially important to realize for MRP supertrees, where the natural temptation is to interpret the resulting branch lengths in terms of the number of synapomorphies supporting that branch. Although the MRP supertree is indeed derived from a parsimony analysis, there is no connection with the original data such that one cannot talk about shared derived characters per se. Instead, meaningful branch lengths for the supertrees have to be obtained by mapping the primary character data a posteriori onto the topology of the supertree (e.g., Song et al. 2012), possibly in combination with calibration data to obtain real divergence-time estimates (e.g., Nyakatura and Bininda-Emonds 2012).

3.4.4 Step 4: Assessing Support Within the Supertree

As pointed out by Purvis (1995b), the use of the nonparametric bootstrap to summarize the nodal support within a supertree was invalid because the inherent non-independence of the additive binary coding (Farris et al. 1970) underlying matrix representation violates a key assumption of the bootstrap. Although this is correct, the real problem with the application of this and any other character-based support method (e.g., Bremer support) is that all fail to account for the fact that the raw data of a supertree analysis are the source trees and not the character data underlying them or even the pseudo-characters derived from them via matrix representation.

Although their development was somewhat delayed and nowhere near as well explored as the creation of new supertree methods, several supertree-specific support measures now exist. One class contains methods that are analogous to the

nonparametric bootstrap for character data (Felsenstein 1985a), except that the source trees are instead resampled with replacement. This procedure has been implemented in the software package CLANN (Creevey and McInerney 2005), but obviously only applies to the supertree methods available within it. An implementation of this method, multilocus bootstrapping, is also available for MLC-based supertree methods (Liu et al. 2010). A variation on this basic scheme, stratified bootstrapping, builds the supertree in each replicate from a randomly chosen tree from the bootstrap profile of each gene tree (Burleigh et al. 2006). Although this point has not been examined, stratified bootstrapping might be able to account in a limited fashion for hidden support within the raw character data as far as it is expressed among the trees in the individual bootstrap profiles.

As with the normal bootstrap, a clear disadvantage of this method in general is its high computational load in that n replicates of the supertree analysis are essentially being performed. Although these searches can be simplified to save time (e.g., performing no branch swapping like in PAUP*'s (Swofford 2002) faststep bootstrap search), this solution invokes other problems because the individual bootstrap trees will not be as optimal, thereby potentially biasing the overall bootstrap frequencies in some unknown manner. Another potential problem with a supertree bootstrap analysis is that some bootstrap replicates might contain nonoverlapping sets of trees and/or might not contain the full taxon set present across all source trees, with this probability increasing as the degree of overlap among the set of source trees decreases. Again, such bootstrap replicates will obtain a completely unresolved supertree, thereby artificially decreasing the overall bootstrap frequencies. Although this scenario is also possible for an equivalent supermatrix analysis (i.e., character partitions that do not overlap with respect to their taxa), it is less likely given the larger number of characters compared to source trees (e.g., 10 partially nonoverlapping source trees might be obtained from 10,000 base pairs worth of sequence data). A potential solution here might be to include a seed tree in each bootstrap replicate, should one be present in the global analysis, to again provide a scaffold ensuring sufficient taxonomic overlap and complete taxon coverage.

Importantly, these supertree bootstrap methods not only provide an estimate of the differential support among the nodes within the supertree, but the profile of bootstrap supertrees is also useful for comparative analyses. Given that the results of the latter are dependent on the accuracy of the underlying phylogenetic tree, accounting for uncertainty/error in the latter is desirable such that the recent trend has been to perform comparative tests on a distribution of trees rather than on a single point estimate of the phylogeny (e.g., Arnold et al. 2010; Jetz et al. 2012; see also Chaps. 10–12). Typically, this distribution is obtained from a Markov chain Bayesian framework; however, there seems to be no reason why a profile of bootstrap trees cannot fulfill the same purpose.

A second class of support measures comprises those that directly quantify the degree of conflict between the supertree and the set of source trees. Examples here include the QS index (Bininda-Emonds 2003) and V (Wilkinson et al. 2005). Compared to the bootstrap, these methods are extremely rapid because both the supertree and the set of source trees have already been computed. Nevertheless, an

inherent difficulty of the method is how to define support versus conflict in the case of missing taxa between the supertree and source tree (Bininda-Emonds 2003). For example, do source trees that contain either taxon A or taxon B, but not both, support or contradict a sister-group relationship between A and B specified by the supertree, or are they uninformative? Both the QS index and V take different approaches to this problem and it is unclear which, if either, is better.

Finally, although supertree methods like PhySIC and PhySIC_IST guarantee that no node on the supertree is contradicted by any of the source trees (Ranwez et al. 2007; Scornavacca et al. 2008), assessing the nodal support on these supertrees using either of the two classes of methods above is arguably still recommended. Key is that both PhySIC and PhySIC_IST do not assure that all source trees directly support a given supertree node, such that while all nodes are not contradicted, some might enjoy more absolute support than others.

3.5 Conclusions

As mentioned in the Introduction (Sect. 3.1), the molecular revolution has arguably been more revolutionary in terms of the massive amounts of phylogenetic data it has provided rather than in the novel hypotheses of phylogenetic relationships it has produced. The latter stability also extends to the gene tree/species tree dichotomy that forms the basis of this chapter, where the reality is that most phylogenetic methods and analytical frameworks seem to be pointing in the same general direction. Thus, the reassuring trend we see is one of growing congruence rather than increasing conflict. Problem areas do remain (e.g., the root of the placental mammals; Teeling and Hedges 2013), but have long been recognized as such, even within any single framework.

Nevertheless, as I have argued in the past (Bininda-Emonds 2004c), a supertree framework—including the MLC model—remains a valid and desirable complement (not alternative) to a pure concatenation-based supermatrix framework, which remains the de facto standard of (molecular) phylogenetics. This point has also been admitted to some extent by even the most vocal critics of supertrees, who minimally see the methodological need for supertrees in piecing together the entire Tree of Life (Gatesy and Springer 2004) and/or do not object to the supertree framework in general (Murphy et al. 2012). More generally, by focusing on different levels of the phylogenetic data set—gene trees versus individual nucleotides, respectively—both the supertree and supermatrix frameworks place slightly different analytical emphases on the same base data set and the use of both approaches in parallel potentially balances out their respective strengths and weaknesses. For example, whereas only a supermatrix framework can account for hidden support, supertrees are better able to account for gene-tree heterogeneity. Given these different foci, analyzing a data set using both frameworks (i.e., essentially parallel partitioned vs. unpartitioned analyses) will therefore provide us with greater confidence in those areas where their results are congruent and greater insight into the causes of any

incongruence where they are not, an approach in agreement with the global-congruence framework of Lapointe et al. (1999). In this way, we will also be better able to establish the frequency of ILS among different taxonomic groups as well as its potential for leading supermatrix-based analyses astray. Moreover, the potential to expand the MLC model in particular to incorporate processes of gene flow and HGT (Liu et al. 2009a) should provide even greater information regarding their frequency and their effects on speciation and phylogenetic history.

Glossary

Hidden support (AKA signal enhancement)	The phenomenon whereby consistent secondary signals among a set of data partitions can overrule their conflicting primary signals to yield a novel solution not to be found among any of the individual data sets. As a simplified example, take the case of two separate gene data sets, each with an aligned length of 1000 nucleotides. In the first data set, 60 % of the positions support a sister-group relationship between A and B (primary signal), whereas 40 % support the clustering of B and C (secondary signal). In the second data set, 60 % support A and C, whereas 40 % support B and C.
	Separate analyses of each data set will yield conflicting results (AB vs. AC); however, when the data sets are combined, each of these solutions is now only supported by 30 % of the data. By contrast, the secondary signals supporting BC are now present among 40 % of the combined data and now form the primary signal. In other words, each separate data set possessed hidden support for BC that could combine and determine the overall solution upon the concatenation of the data sets. Because supertree analyses work with trees as their primary data source, these secondary signals in the raw character data are normally invisible and cannot be accounted for.
Long-branch attraction	An artifact in the phylogenetic analysis of DNA sequence data that was first exposed by Felsenstein (1978) and is a result of SATURATION in such data. Felsenstein observed that taxa at the ends of very long branches that themselves were separated by a short intervening branch often clustered to form sister taxa in a maximum parsimony analysis. Optimization criteria that used an explicit model of

	evolution like maximum likelihood were more immune to this problem. This artifactual attraction of the long branches arises because the taxa are characterized by high rates of molecular evolution (as indicated by the long branches) and concomitant large number of shared convergent changes that, through their high number, are falsely interpreted as evidence for shared common ancestry. It is now known that long-branch attraction is a general problem (i.e., it can affect nonmolecular data, although is far less likely to do so) and can occur even if the branches occur on distant parts of the tree (see Bergsten 2005).
Matrix representation	A long-standing mathematical principle (Ponstein 1966) showing that there is a one-to-one correspondence between a tree (a "directed acyclic graph") and its encoding as a binary matrix. Whereas additive binary coding (Farris et al. 1970) of the tree will derive the matrix, the tree can be recreated from the matrix via analysis of the latter using virtually any optimization criterion (see Fig. 3.2).
NP-complete	A class of nondeterministic polynomial (NP) time methods for which no efficient solution is known and for which the running time increases tremendously with the size of the problem. As such, heuristic rather than exact algorithms must be used beyond a certain problem size, meaning that there is no guarantee that the optimal solution has been found. In phylogenetics, classic examples of NP-complete algorithms include maximum parsimony and maximum likelihood.
Polynomial time	Polynomial time algorithms are said to be "fast" in the sense that they have an efficient solution that scales "reasonably" with the size of the problem. A cogent example here is neighbor joining (NJ), the running time of which scales no worse than the cube of the number of taxa (i.e., $O(n^3)$). This is in stark contrast to the NP-COMPLETE maximum parsimony and maximum-likelihood methods, where the running times scale super-exponentially with respect to the problem size.
Saturation	A phenomenon attributed primarily to DNA sequence data and which arises because of the limited character state space for such data (i.e., the four nucleotides A, C, G, and T). As such, the potential for homoplasy in the form of either convergence or back mutation is high (e.g., two

completely random DNA sequences are expected to be 25 % similar). Saturation, however, can also occur, but is less likely, for both amino-acid and morphological character data.

In practice, saturation is visualized by the degree of divergence between two sequences leveling off or plateauing with time since their divergence because faster evolving sites have experienced multiple substitutions ("multiple hits") with the increased potential for homoplastic similarity. Another method is to examine for deviations from an expected transition: transversion ratio of 1:2 in neutral/silent sites, given the faster rate of evolution for transitions compared to transversions and, again, greater opportunity for multiple hits.

References

Adams EM III (1972) Consensus techniques and the comparison of taxonomic trees. Syst Zool 21:390–397

Adams EM III (1986) N-trees as nestings: complexity, similarity, and consensus. J Classif 3:299–317

Arnold CL, Matthews J, Nunn CL (2010) The 10k Trees website: a new online resource for primate phylogeny. Evol Anthropol 19:114–118

Asher RJ, Müller J (2012) Molecular tools in palaeobiology: divergence and mechanisms. In: Asher RJ, Müller J (eds) From clone to bone: the synergy of morphological and molecular tools in palaeobiology. Cambridge studies in morphology and molecules: new paradigms in evolutionary biology, vol 4. Cambridge University Press, Cambridge, pp 1–15

Baker RH, DeSalle R (1997) Multiple sources of character information and the phylogeny of Hawaiian drosophilids. Syst Biol 46:654–673

Baum BR (1992) Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. Taxon 41:3–10

Beddard FE (1900) A book of whales. G.P. Putnam's Sons, New York

Bergsten J (2005) A review of long-branch attraction. Cladistics 21(2):163–193

Berry V, Bininda-Emonds ORP, Semple C (2013) Amalgamating source trees with different taxonomic levels. Syst Biol 62(2):231–249

Bininda-Emonds ORP (2003) Novel versus unsupported clades: assessing the qualitative support for clades in MRP supertrees. Syst Biol 52(6):839–848

Bininda-Emonds ORP (2004a) The evolution of supertrees. Trends Ecol Evol 19(6):315–322

Bininda-Emonds ORP (2004b) New uses for old phylogenies: an introduction to the volume. In: Bininda-Emonds ORP (ed) Phylogenetic supertrees: combining information to reveal the tree of life, computational biology, vol 4. Kluwer Academic, Dordrecht, the Netherlands, pp 3–14

Bininda-Emonds ORP (2004c) Trees versus characters and the supertree/supermatrix "paradox". Syst Biol 53(2):356–359

Bininda-Emonds ORP (2010) The future of supertrees: bridging the gap with supermatrices. Palaeodiversity 3(Suppl.):99–106

Bininda-Emonds ORP (2011) Inferring the Tree of Life: chopping a phylogenomic problem down to size? BMC Biol 9:59

Bininda-Emonds ORP, Beck RMD, Purvis A (2005) Getting to the roots of matrix representation. Syst Biol 54(4):668–672

Bininda-Emonds ORP, Bryant HN (1998) Properties of matrix representation with parsimony analyses. Syst Biol 47(3):497–508

Bininda-Emonds ORP, Jones KE, Price SA, Cardillo M, Grenyer R, Purvis A (2004) Garbage in, garbage out: data issues in supertree construction. In: Bininda-Emonds ORP (ed) Phylogenetic supertrees: combining information to reveal the Tree of Life, computational biology, vol 4. Kluwer Academic, Dordrecht, the Netherlands, pp 267–280

Bininda-Emonds ORP, Jones KE, Price SA, Grenyer R, Cardillo M, Habib M, Purvis A, Gittleman JL (2003) Supertrees are a necessary not-so-evil: a comment on Gatesy et al. Syst Biol 52 (5):724–729

Bininda-Emonds ORP, Sanderson MJ (2001) Assessment of the accuracy of matrix representation with parsimony supertree construction. Syst Biol 50(4):565–579

Bremer K (1988) The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. Evolution 42:795–803

Bruen TC, Bryant D (2008) Parsimony via consensus. Syst Biol 57(2):251–256

Burleigh JG, Driskell AC, Sanderson MJ (2006) Supertree bootstrapping methods for assessing phylogenetic variation among genes in genome-scale data sets. Syst Biol 55(3):426–440

Chaudhary R, Bansal MS, Wehe A, Fernandez-Baca D, Eulenstein O (2010) iGTP: a software package for large-scale gene tree parsimony analysis. BMC Bioinformatics 11:574

Chen D, Diao L, Eulenstein O, Fernández-Baca D, Sanderson MJ (2003) Flipping: a supertree construction method. In: Janowitz MF, Lapointe F-J, McMorris FR, Mirkin B, Roberts FS (eds) Bioconsensus, vol 61., DIMACS Series in discrete mathematics and theoretical computer scienceAmerican Mathematical Society, Providence, RI, pp 135–160

Chen D, Eulenstein O, Fernández-Baca D (2004) Rainbow: a toolbox for phylogenetic supertree construction and analysis. Bioinformatics 20(16):2872–2873

Chippindale PT, Wiens JJ (1994) Weighting, partitioning, and combining characters in phylogenetic analysis. Syst Biol 43:278–287

Cotton JA, Page RDM (2004) Tangled trees from multiple markers: reconciling conflict between phylogenies to build molecular supertrees. In: Bininda-Emonds ORP (ed) Phylogenetic supertrees: combining information to reveal the tree of life, computational biology, vol 4. Kluwer Academic, Dordrecht, the Netherlands, pp 107–125

Cotton JA, Wilkinson M (2007) Majority-rule supertrees. Syst Biol 56(3):445–452

Creevey CJ, McInerney JO (2005) Clann: investigating phylogenetic information through supertree analyses. Bioinformatics 21(3):390–392

Davis KE, Hill J (2010) The supertree tool kit. BMC Res Notes 3:95

de Queiroz A, Donoghue MJ, Kim J (1995) Separate versus combined analysis of phylogenetic evidence. Annu Rev Ecol Syst 26:657–681

Degnan JH, Rosenberg NA (2006) Discordance of species trees with their most likely gene trees. PLoS Genet 2(5):762–768

Edwards SV (2009) Is a new and general theory of molecular systematics emerging? Evolution 63(1):1–19

Farris JS, Kluge AG, Eckhardt MJ (1970) A numerical approach to phylogenetic systematics. Syst Zool 19:172–191

Felsenstein J (1978) Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool 27:401–410

Felsenstein J (1985a) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791

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

Gatesy J, Matthee C, DeSalle R, Hayashi C (2002) Resolution of a supertree/supermatrix paradox. Syst Biol 51(4):652–664

Gatesy J, O'Grady P, Baker RH (1999) Corroboration among data sets in simultaneous analysis: hidden support for phylogenetic relationships among higher level artiodactyl taxa. Cladistics 15(3):271–313

Gatesy J, Springer MS (2004) A critique of matrix representation with parsimony supertrees. In: Bininda-Emonds ORP (ed) Phylogenetic supertrees: combining information to reveal the tree of life, computational biology, vol 4. Kluwer Academic, Dordrecht, the Netherlands, pp 369–388

Gatesy J, Springer MS (2013) Concatenation versus coalescence versus "concatalescence". Proc Natl Acad Sci U S A 110(13):E1179–E1179

Goodman M, Czelusniak J, Moore GW, Romero-Herrera AE, Matsuda G (1979) Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Syst Zool 28(2):132–163

Gordon AD (1986) Consensus supertrees: the synthesis of rooted trees containing overlapping sets of labeled leaves. J Classif 3:31–39

Graybeal A (1998) Is it better to add taxa or characters to a difficult phylogenetic problem? Syst Biol 47(1):9–17

Hailer F, Kutschera VE, Hallstrom BM, Klassert D, Fain SR, Leonard JA, Arnason U, Janke A (2012) Nuclear genomic sequences reveal that polar bears are an old and distinct bear lineage. Science 336(6079):344–347. doi:10.1126/science.1216424

Harvey PH, Pagel MD (1991) The comparative method in evolutionary biology. Oxford University Press, Oxford

Hillis DM (1987) Molecular versus morphological approaches to systematics. Annu Rev Ecol Syst 18:23–42

Jetz W, Thomas GH, Joy JB, Hartmann K, Mooers AO (2012) The global diversity of birds in space and time. Nature 491:444–448

Kluge AG (1989) A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). Syst Zool 38:7–25

Lanfear R, Calcott B, Ho SYW, Guindon S (2012) PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol Biol Evol 29(6):1695–1701

Lapointe F-J, Cucumel G (1997) The average consensus procedure: combination of weighted trees containing identical or overlapping sets of taxa. Syst Biol 46(2):306–312

Lapointe F-J, Kirsch JAW, Hutcheon JM (1999) Total evidence, consensus, and bat phylogeny: a distance based approach. Mol Phylogenet Evol 11(1):55–66

Lapointe F-J, Levasseur C (2004) Everything you always wanted to know about the average consensus, and more. In: Bininda-Emonds ORP (ed) Phylogenetic supertrees: combining information to reveal the tree of life, computational biology, vol 4. Kluwer Academic, Dordrecht, the Netherlands, pp 87–105

Lee MS, Camens AB (2009) Strong morphological support for the molecular evolutionary tree of placental mammals. J Evol Biol 22 (11):2243–2257. doi:JEB1843 [pii] 10.1111/j.1420-9101.2009.01843.x

Lindqvist C, Schuster SC, Sun Y, Talbot SL, Qi J, Ratan A, Tomsho LP, Kasson L, Zeyl E, Aars J, Miller W, Ingolfsson O, Bachmann L, Wiig O (2010) Complete mitochondrial genome of a Pleistocene jawbone unveils the origin of polar bear. Proc Natl Acad Sci U S A 107(11):5053–5057. doi:10.1073/pnas.0914266107

Liu F-GR, Miyamoto MM, Freire NP, Ong PQ, Tennant MR, Young TS, Gugel KF (2001) Molecular and morphological supertrees for eutherian (placental) mammals. Science 291:1786–1789

Liu L, Yu L (2010) Phybase: an R package for species tree analysis. Bioinformatics 26:962–963

Liu L, Yu LL, Kubatko L, Pearl DK, Edwards SV (2009a) Coalescent methods for estimating phylogenetic trees. Mol Phylogenet Evol 53(1):320–328

Liu L, Yu LL, Pearl DK, Edwards SV (2009b) Estimating species phylogenies using coalescence times among sequences. Syst Biol 58(5):468–477

Liu LA, Yu LL, Edwards SV (2010) A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol Biol 10

Maddison WP (1997) Gene trees in species trees. Syst Biol 46(3):523–536

Miller W, Schuster SC, Welch AJ, Ratan A, Bedoya-Reina OC, Zhao F, Kim HL, Burhans RC, Drautz DI, Wittekindt NE, Tomsho LP, Ibarra-Laclette E, Herrera-Estrella L, Peacock E, Farley S, Sage GK, Rode K, Obbard M, Montiel R, Bachmann L, Ingolfsson O, Aars J, Mailund T, Wiig O, Talbot SL, Lindqvist C (2012) Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. Proc Natl Acad Sci USA 109(36):E2382–E2390. doi:10.1073/pnas.1210506109

Mossel E, Roch S (2007) Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. http://arxiv.org/abs/0710.0262

Murphy WJ, Janecka JE, Stadler T, Eizirik E, Ryder OA, Gatesy J, Meredith RW, Springer MS (2012) Response to comment on "impacts of the cretaceous terrestrial revolution and KPg extinction on mammal diversification". Science 337(6090):34

Nguyen N, Mirarab S, Warnow T (2012) MRL and SuperFine plus MRL: new supertree methods. Algorithms Mol Biol 7(1):3

Nyakatura K, Bininda-Emonds ORP (2012) Updating the evolutionary history of Carnivora (mammalia): a new species-level supertree complete with divergence time estimates. BMC Biol 10:12

Page RDM (1994) Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. Syst Biol 43(1):58–77

Page RDM (2002) Modified mincut supertrees. In: Guigó R, Gusfield D (eds) Proceedings of Algorithms in bioinformatics, second international workshop, WABI, Rome, Italy. Lecture Notes in computer science, vol 2452. Springer, Berlin, pp 537–552, 17–21 Sept 2002

Piaggio-Talice R, Burleigh JG, Eulenstein O (2004) Quartet supertrees. In: Bininda-Emonds ORP (ed) Phylogenetic supertrees: combining information to reveal the tree of life, computational biology, vol 4. Kluwer Academic, Dordrecht, the Netherlands, pp 173–191

Ponstein J (1966) Matrices in graph and network theory. Van Gorcum, Assen, Netherlands

Purvis A (1995a) A composite estimate of primate phylogeny. Philos Trans R Soc Lond B 348:405–421

Purvis A (1995b) A modification to Baum and Ragan's method for combining phylogenetic trees. Syst Biol 44:251–255

Ragan MA (1992) Phylogenetic inference based on matrix representation of trees. Mol Phylogenet Evol 1:53–58

Rannala B, Yang ZH (2003) Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164:1645–1656

Ranwez V, Berry V, Criscuolo A, Fabre PH, Guillemot S, Scornavacca C, Douzery EJ (2007) PhySIC: a veto supertree method with desirable properties. Syst Biol 56 (5):798–817. doi:782748826 [pii] 10.1080/10635150701639754

Rokas A, Holland PW (2000) Rare genomic changes as a tool for phylogenetics. Trends Ecol Evol 15(11):454–459

Ronquist F (1996) Matrix representation of trees, redundancy, and weighting. Syst Biol 45:247–253

Ronquist F, Huelsenbeck JP, Britton T (2004) Bayesian supertrees. In: Bininda-Emonds ORP (ed) Phylogenetic supertrees: combining information to reveal the tree of life, computational biology, vol 4. Kluwer Academic, Dordrecht, the Netherlands, pp 193–224

Rosenberg NA (2013) Discordance of species trees with their most likely gene trees: a unifying principle. Mol Biol Evol 30(12):2709–2713. doi:10.1093/molbev/mst160

Roshan U, Moret BME, Williams TL, Warnow T (2004) Performance of supertree methods on various data set decompositions. In: Bininda-Emonds ORP (ed) Phylogenetic supertrees: combining information to reveal the tree of life, computational biology, vol 4. Kluwer Academic, Dordrecht, the Netherlands, pp 301–328

Ross HA, Rodrigo AG (2004) An assessment of matrix representation with compatibility in supertree construction. In: Bininda-Emonds ORP (ed) Phylogenetic supertrees: combining information to reveal the tree of life, computational biology, vol 4. Kluwer Academic, Dordrecht, the Netherlands, pp 35–63

Salamin N, Hodkinson TR, Savolainen V (2002) Building supertrees: an empirical assessment using the grass family (Poaceae). Syst Biol 51(1):136–150

Sanderson MJ, Donoghue MJ, Piel W, Eriksson T (1994) TreeBASE: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. Am J Bot 81(6):183

Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18(3):502–504

Scornavacca C, Berry V, Lefort V, Douzery EJ, Ranwez V (2008) PhySIC_IST: cleaning source trees to infer more informative supertrees. BMC Bioinformatics 9:413. doi:1471-2105-9-413 [pii] 10.1186/1471-2105-9-413

Semple C, Steel M (2000) A supertree method for rooted trees. Discrete Appl Math 105(1–3):147–158

Smith SA, Beaulieu JM, Stamatakis A, Donoghue MJ (2011) Understanding angiosperm diversification using small and large phylogenetic trees. Am J Bot 98(3):404–414

Song S, Liu L, Edwards SV, Wu SY (2012) Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proc Natl Acad Sci USA 109(37):14942–14947

Stamatakis A (in press) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. doi:10.1093/bioinformatics/btu033

Steel M, Dress AWM, Böcker S (2000) Simple but fundamental limitations on supertree and consensus tree methods. Syst Biol 49(2):363–368

Steel M, Rodrigo A (2008) Maximum likelihood supertrees. Syst Biol 57(2):243–250

Strimmer K, von Haeseler A (1996) Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. Mol Biol Evol 13:964–969

Swenson MS, Suri R, Linder CR, Warnow T (2012) SuperFine: fast and accurate supertree estimation. Syst Biol 61(2):214–227

Swofford DL (2002) PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, Massachusetts

Teeling EC, Hedges SB (2013) Making the impossible possible: rooting the tree of placental mammals. Mol Biol Evol 30(9):1999–2000

Thorley JL, Wilkinson M (2003) A view of supertree methods. In: Janowitz MF, Lapointe F-J, McMorris FR, Mirkin B, Roberts FS (eds) Bioconsensus, vol 61., DIMACS series in discrete mathematics and theoretical computer scienceAmerican Mathematical Society, Providence, RI, pp 185–193

Wehe A, Bansal MS, Burleigh JG, Eulenstein O (2008) DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. Bioinformatics 24(13):1540–1541

Wilkinson M, Pisani D, Cotton JA, Corfe I (2005) Measuring support and finding unsupported relationships in supertrees. Syst Biol 54(5):823–831

Wilkinson M, Thorley JL, Pisani D, Lapointe F-J, McInerney JO (2004) Some desiderata for liberal supertrees. In: Bininda-Emonds ORP (ed) Phylogenetic supertrees: combining information to reveal the tree of life, computational biology, vol 4. Kluwer Academic, Dordrecht, the Netherlands, pp 227–246

Wu SY, Song S, Liu L, Edwards SV (2013) Reply to Gatesy and Springer: The multispecies coalescent model can effectively handle recombination and gene tree heterogeneity. Proc Natl Acad Sci USA 110(13):E1180–E1180

Chapter 4
Graphical Methods for Visualizing Comparative Data on Phylogenies

Liam J. Revell

Abstract Phylogenies have emerged as central in evolutionary biology over the past three decades or more, and an extraordinary expansion in the breadth and sophistication of phylogenetic comparative methods has played a large role in this growth. In this chapter, I focus on a somewhat neglected area: the use of graphical methods to simultaneously represent comparative data and trees. As this research area is theoretically very broad, I have concentrated on new methods developed by me, or techniques devised by others and implemented by me as part of my R phylogenetics package, phytools. I describe a variety of methods in this chapter, including approaches that can be used to map reconstructed discrete or continuous character evolution on trees; techniques for projecting phylogenetic trees into morphospace; and methods for visualizing phylogenies in the context of a global or regional geographic map. In this chapter, my intention is not merely to showcase new methods that I have developed. Rather, I have also dedicated considerable attention to detailing the algorithms and computational techniques required for these approaches with the hope that this chapter will become a resource or jumping-off point for researchers interested in building new, more advanced approaches and methods in this area.

4.1 Introduction

No one would seriously dispute the contention that a well-designed and informative figure can replace at least a thousand words, if not more, in a contemporary scientific publication. Visualization can also play an integral role in the preliminary analysis of new data and in generating new hypotheses which can be

L. J. Revell (✉)
Department of Biology, University of Massachusetts Boston, Boston, MA 02125, USA
e-mail: liam.revell@umb.edu

explored with more rigorous tests. However, effective graphical methods for new types of data and analysis in phylogenetic comparative biology may also require some creative new method development for visualization (e.g., Sidlauskas 2008; Revell 2013). In this chapter, I'm going to describe and illustrate several new approaches—devised by myself or others and implemented in my R package, phytools (Revell 2012)—for visualizing comparative data on phylogenies. Specifically, I'll focus on visualization methods that can simultaneously show the phylogenetic tree and a set of comparative data for discrete and continuously valued phenotypic traits.

Large evolutionary changes take place over thousands of generations to millions of years. In many cases, phylogenetic comparative biology—the theory and practice of drawing evolutionary inferences from phylogenies and comparative data for phenotypic characters—represents our best or only recourse for studying evolution on these vast timescales (Felsenstein 1985, 1988; Harvey and Pagel 1991; Mahler et al. 2010; Nunn 2011). Phylogenetic comparative methods have advanced considerably in recent years (e.g., Butler and King 2004; O'Meara et al. 2006; Bokma 2008; Fitzjohn 2010; Eastman et al. 2011; Felsenstein 2012; Revell et al. 2012; Beaulieu et al. 2013; Revell 2014; reviewed in Glor 2010; O'Meara 2012; Pennell and Harmon 2013). Many of the chapters of this book exemplify these great strides. However, in some cases, these new methods and new types of data for comparative biology also present us with new challenges in visualization. Specifically, the most efficient, visually appealing, and informative way to simultaneously represent phylogenetic and phenotypic information in a single plot is not always clear.

Since a simple plot of the phylogeny forms the basis for several of the visualization methods that I'll describe in this chapter, I'm going to begin (in Sect. 4.2, below) by detailing the general algorithm that can be used to draw two common types of tree plots. In subsequent sections, I'll focus my attention more specifically on the challenges of simultaneously visualizing phylogenetic relationships and trait data for phenotypic characters. In Sect. 4.3, I'll concentrate on discrete character methods. I'll describe the comparative method called stochastic character mapping (Nielsen 2002; Huelsenbeck et al. 2003; Bollback 2006) and illustrate how a single stochastic map can be plotted on the branches and nodes of a phylogeny (Sect. 4.3.1). Next, I'll detail and illustrate two different approaches for aggregating the results of many stochastic mappings (Sects. 4.3.2 and 4.3.3; Revell 2013). Then, in Sect. 4.4, I'll move on to several different methods that have been developed for continuously valued phenotypic traits. The first and second methods (described in Sects. 4.4.1 and 4.4.2) involve some kind of projection of the phylogeny into a space that is either fully or partially defined by our phenotypic trait data in two or three dimensions (e.g., Sidlauskas 2008; Evans et al. 2009). The third method (described in Sect. 4.4.3) involves directly mapping the reconstructed evolution of a continuous trait onto the branches of a plotted tree. I also show how we can combine both types of plots to create a "phylogenetic scatterplot matrix" suitable for multidimensional continuous trait data. In Sect. 4.5, I'll describe a few additional new approaches, including the projection

of a tree onto a geographic map, and the combination of discrete and continuous character methods into a single plot. In Sect. 4.6, I'll give a brief introduction to the technical matter of programming phylogeny plotting methods in R (R Core Team 2013). Finally, in Sect. 4.7, I'll try to provide some concluding thoughts on the challenges of visualizing phylogenetic and comparative data, the "paper paradigm" (Rosindell and Harmon 2012), and some possible future developments of this field.

Before I begin, I should emphasize again that this chapter is not intended as a comprehensive survey of phylogenetic visualization methods, nor even of visualization methods for phylogenetic comparative biology. Rather, I have focused specifically on methods that I have worked on in some way. By providing detail on these methods, rather than a superficial survey of all approaches, I hope to inspire readers of this chapter to think about novel techniques that are suitable for their (idiosyncratic or general) problem or data. Hopefully, the content of this chapter can become a starting point for additional methodological innovation and discovery by other researchers.

Although all the methods of this chapter are implemented in my phytools R package, plotting methods in phytools make extensive use of R base graphics (R Core Team 2013), as well as some other packages such as scatterplot3d (Ligges and Mächler 2003), maps (Becker et al. 2013), plotrix (Lemon 2006), and rgl (Adler and Murdoch 2013). In addition, phytools depends internally on ape (Paradis et al. 2004) and phangorn (Schliep 2011) for their extensive suite of functions for reading, writing, manipulating, and analyzing phylogenetic trees.

4.2 The General Problem of Drawing Trees

In this section, I'll briefly describe the basic general algorithm for taking a tree stored in computer memory (or, hypothetically, in your own memory) and drawing that tree onto a piece of paper or a plotting object in R. Since there are already many different tools for tree drawing available, this section will primarily appeal to researchers interested in programming new visualization methods for phylogenies (or understanding how existing methods are programmed). Readers that are not interested in such things can probably skip this section. Here, I'm going to focus on the general algorithms for tree plotting, which would apply equally to any programming language or development environment; however, in Sect. 4.6, toward the end of this chapter, I'm going to go on to provide some specific code in R to replicate one of these algorithms. Here, I'm going to concentrate on two different types of trees that are also among the easiest to draw: square and circular phylograms with intermediate node placement (Felsenstein 2004). Examples of these two tree-plotting methods are illustrated in Fig. 4.1a and b, respectively. I have to presume that these algorithms have previously been independently discovered by anyone who has ever programmed a tree-plotting function; however,
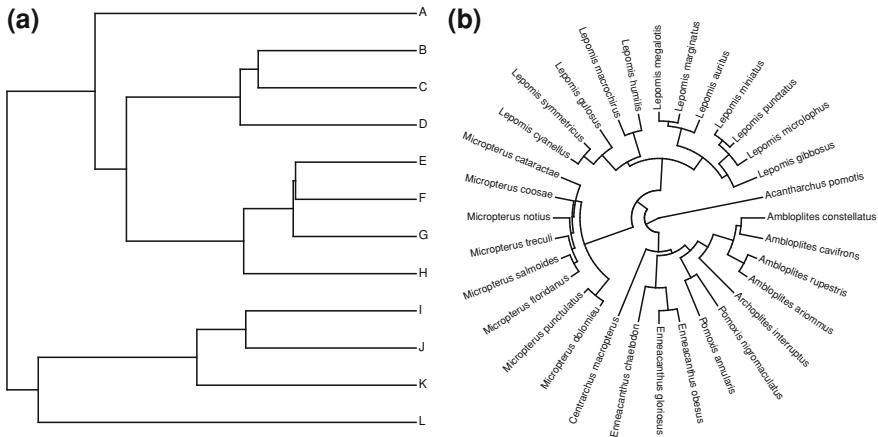
Fig. 4.1 **a** A square phylogram representing a simulated random tree. **b** A circular phylogram showing the empirical phylogeny of Centrarchidae [from Near et al. (2005)]

I'm not aware (and I could be wrong, of course) of them having been written down—at least not in the phylogeny literature.

Figure 4.1a shows a stochastic, pure-birth tree (i.e., a "Yule tree") plotted as a rightward square phylogram with intermediate node placement (Felsenstein 2004). *Rightward* refers to the orientation of edges (when taken as vectors leading from parent to daughter); *phylogram* just means that the plotted edges are proportional in length to the branches of the tree; and *intermediate node placement* refers to the (in this case) vertical position of ancestral nodes—in other words, we have positioned them vertically intermediately between the uppermost and lowermost daughter nodes. Note that although the algorithm is described specifically for a rightward orientation (in other words a phylogeny that "grows" from left to right on our page; Felsenstein 2004), changing to a leftward orientation, or an upward or downward orientation, simply requires that we change the sign of x, or flip x and y, or do first one and then the other.

To create a graph in this style, the first step (step 1) is assigning vertical positions to all of the tips in the tree. To do this, we first have to sort the tips into what I'm going to refer to as *cladewise* order (Paradis 2012). This just means that tips in a clade are adjacent to each other in the ordering. In the case of Fig. 4.1, this means that A, B, C, D, E, F, G, H, I, J, K, L; B, C, D, E, F, G, H, A, I, J, K, L; or E, F, G, H, B, C, D, A, I, J, K, L are all valid cladewise orderings of the twelve taxa in our tree. (There are also many other valid orderings.)

If this ordering seems like it could be complicated to obtain, then it might be helpful to note that a left-to-right (or right-to-left) reading of the tip labels in a Newick style tree is guaranteed to produce tip labels in cladewise order. For example, the Newick strings below:

((A:0.78,(((B:0.39,C:0.39):0.04,D:0.43):0.28,(((E:0.29,F:0.29):0.01,G:0.3):0.12,
H:0.42):0.29):0.08):0.22,(((I:0.42,J:0.42):0.12,K:0.54):0.39,L:0.92):0.08);,
(((((B:0.39,C:0.39):0.04,D:0.43):0.28,(((E:0.29,F:0.29):0.01,G:0.3):0.12,H:0.42):
0.29):0.08,A:0.78):0.22,(((I:0.42,J:0.42):0.12,K:0.54):0.39,L:0.92):0.08);, and
((((((E:0.29,F:0.29):0.01,G:0.3):0.12,H:0.42):0.29,((B:0.39,C:0.39):0.04,D:0.43):
0.28):0.08,A:0.78):0.22,(((I:0.42,J:0.42):0.12,K:0.54):0.39,L:0.92):0.08);

are all valid representations of the tree in Fig. 4.1a, and thus, all contain tip labels (read left to right or right to left) in a cladewise ordering. Newick tree format is the most widely used way to record phylogenies in plain text. Newick format uses parentheses, commas, and colons to represent hierarchical and sister relationships, and branch lengths, respectively (Archie et al. 1986). For more information about the Newick format, readers should refer to Felsenstein (2004).

Having ordered the tips in this way, we can now go ahead and assign each tip a vertical position evenly spaced from 1 through n for n tips (step 2). This step is shown in Fig. 4.2a. (We could have just as well assigned values from n through 1, $-n/2$ through $n/2$, 100 through $100 \times n$, or 0.1 through $0.1 \times n$, etc., so long as we are prepared to resize the vertical axis of our plotting area accordingly.)

Now, we conduct a post-order traversal of the tree. This means we descend from the tips to the root of the tree passing through each daughter node before its parent. At each node, we can assign a vertical position to the node (and its preceding edge) that is intermediate between the two daughters (step 3). This step is illustrated in Fig. 4.2b, in which the horizontal dashed line indicating the vertical position of each internal node is labeled with a list of the tips descended from that node. If there are more than two daughters, in other words, if the node contains a multifurcation, we compute the average of the lowermost and the uppermost daughters (Felsenstein 2004). Having done this for all internal nodes, we are now in possession of the vertical position of all plotted edges in the tree.

The next step (step 4) is to compute the horizontal starting and ending points of each branch in the tree. To do this, we start at the root and use a pre-order tree traversal, which means that we traverse each parent node before its daughters. As we traverse the tree up from the root, we compute the starting horizontal position of an edge as the sum of all preceding branch lengths in the path from the root to the parent (starting) node of that edge. The ending point of the same edge is merely this value plus the branch length of the current edge. The points computed in this step are shown in Fig. 4.2c. Now, we have the vertical positions (from step 3) and the starting and ending points of each branch in the tree. When we plot the horizontal lines that connect these points, we've plotted all the branches in our phylogeny in their correct horizontal and vertical positions. This step is shown in Fig. 4.2d.

To add the relationships between species and clades (step 5), in other words, the vertical lines in our plot of Fig. 4.1a, we start by taking each internal node in the tree including the root. We go to its height on the horizontal axis, and we draw a vertical line connecting the uppermost and lowermost vertical positions of its two or more daughters. This step is illustrated in Fig. 4.2e. Having ordered the tip taxa
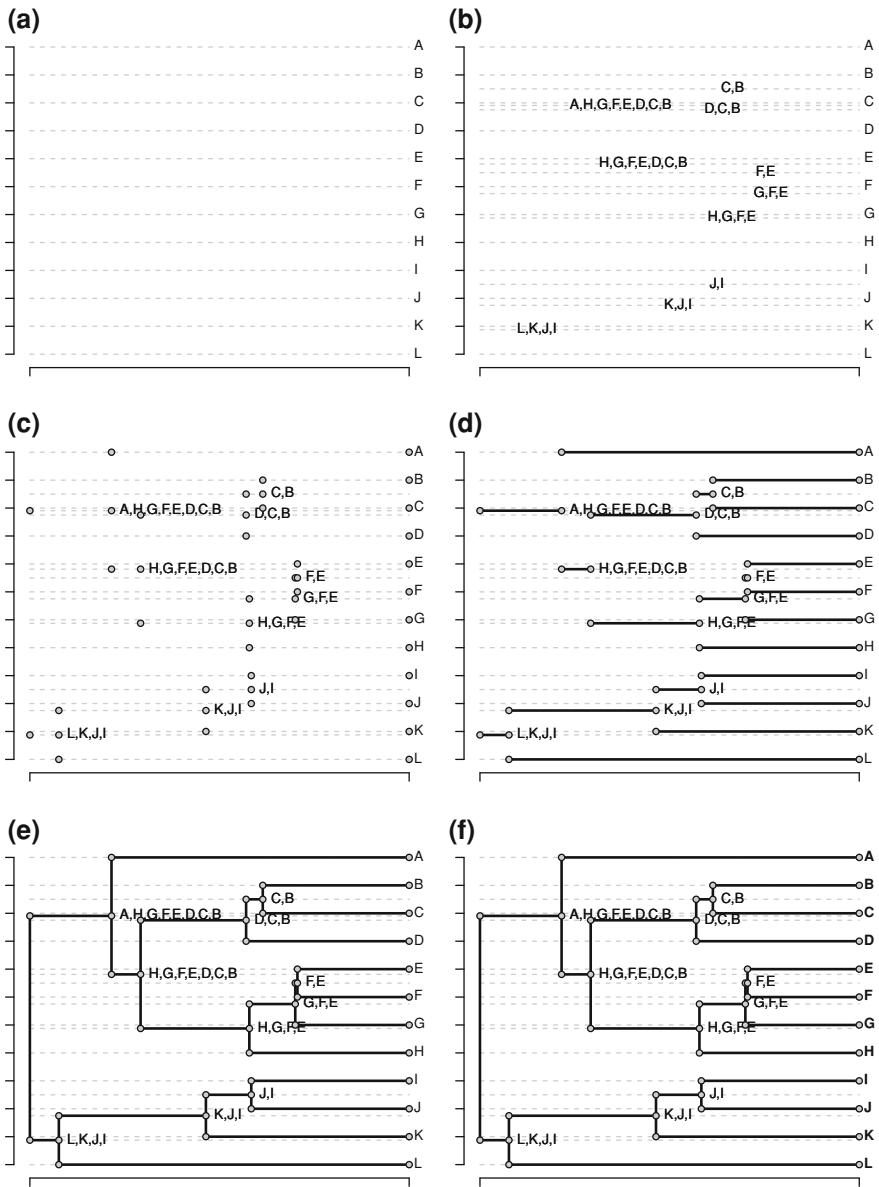
Fig. 4.2 An illustration of the algorithm for drawing a rightward square phylogram with intermediate node placement using the simulated tree of Fig. 4.1a. **a** Order the tips of the tree in "cladewise" order. **b** Conduct a post-order traversal of the tree and compute the vertical position of internal edges as the average of the highest and lowest daughter edges. The vertical positions of terminal and internal edges of the tree are shown as horizontal dashed lines in panels (**a**) and (**b**). **c** Conduct a pre-order tree traversal and record the height above the root of the starting and ending points of each edge. **d** Draw edges. **e** Plot the relationships between edges by going to each internal node and adding a vertical line connecting the highest and lowest daughter edges. **f** Add tip labels at the end of each terminal edge

by clade before assigning their vertical positions in step 2, we've guaranteed that our tree is "untangled"; in other words, no vertical lines cross any of our horizontally drawn edges. Finally, to include labels (step 6), we merely add text to our plot for taxa 1 through n at each of the 1 through n vertical positions we assigned in step 2, in this case using the horizontal positions of the end of each terminal edge, computed in step 4. This final step is illustrated in Fig. 4.2f.

For circular phylograms, such as the phylogeny of centrarchid fishes from Near et al. (2005) shown in Fig. 4.1b, we conduct steps 1 through 4, just as described above for rightward square phylograms. Then, however, we translate our node heights (above the root, two for each edge) and vertical edge positions to the coordinate system of our circular plot using the formulae $\mathbf{x}_i = \mathbf{r}_i \cdot \cos(Y_i)$ and $\mathbf{y}_i = \mathbf{r}_i \cdot \sin(Y_i)$, where $\mathbf{r}_i$ (radius) is the set of heights above the root for edge i and Y_i is its vertical position in the square tree. We connect each parent and daughter subtending edge i using a radial line from $(x_{1,i}, y_{1,i})$ to $(x_{2,i}, y_{2,i})$. Then, at each internal node, j, we draw an arc of radius r_j spanning the lowermost to the uppermost daughter edges of j. Finally, we plot our 1 through n labels at the end of each terminal edge. Normally, we would angle the labels using the same angle as the terminal edge, but then flip the orientation of the label by $180°$ for labels plotted between $90°$ and $270°$ from horizontal (e.g., see Fig. 4.1b). Circular or fan-style plots provide the advantage of allowing larger phylogenies to be represented in the same plotting area (e.g., Edwards et al. 2010), sometimes even with readable tip labels; however, they create the disadvantage that because time since the root is represented by radial distance from the origin (rather than horizontal distance), contemporaneous nodes and tips are a little bit more difficult to identify.

4.3 Discrete Character Methods

4.3.1 Mapping a Single Discrete Character on the Tree

The first plotting method for comparative data that I am going to describe is the relatively simple approach of visualizing the reconstructed history of a discretely valued character trait obtained from a phylogenetic method called stochastic character mapping (Nielsen 2002; Huelsenbeck et al. 2003). Stochastic mapping is a procedure in which we randomly sample possible character histories for a discrete trait such that the probability of sampling any specific history varies in direct relation to its posterior probability under our model of trait evolution (generally, a continuous-time discrete-state Markov chain), given our tree and data. Stochastic character mapping is described in more detail by Nielsen (2002), Huelsenbeck et al. (2003), and Bollback (2006).

Briefly, to generate a single stochastic character map on the tree, we first sample a joint reconstruction of our discrete character across all the nodes of the tree conditioned on an instantaneous transition matrix between states, $\mathbf{Q}$, and our

discrete character data, **x**. (Where $\boldsymbol{Q}$ comes from we will leave aside for the moment.) These states are sampled from their joint posterior probability distribution following Bollback (2006). Next, we simulate changes along the edges of the tree using a rejection procedure. We obtain the waiting times for changes between states by drawing randomly from an exponential distribution with rate $-Q_{ii}$, given initial state i. If the time is shorter than the total branch length of the current edge, we simulate another change, and then another, and so on, until we reach the end of the branch. At each change, we determine the new state by picking a state at random with probability $\Pr(j) = Q_{ij} \Big/ \sum_{j=1}^{n, j \neq i} Q_{ij}$, for any derived state j. Here, n is the total number of different states for our discrete character. If the starting and ending states for the branch match our stochastic joint sampled node states, we have successfully simulated a stochastic history for that branch. If not, then we reject our simulation and repeat it until we have sampled a history with starting and ending states that agree with our stochastically sampled states for the nodes subtending that branch. Stochastic mapping is implemented in phytools (Revell 2012).

There are two different procedures that we can use to obtain our continuous-time discrete-state Markov chain transition matrix, $\boldsymbol{Q}$. We can sample $\boldsymbol{Q}$ using Bayesian MCMC, which I'll refer to as the full hierarchical Bayesian approach; or we can fix $\boldsymbol{Q}$ at its most likely value, which I'm going to call an empirical Bayesian approach (e.g., Yang 2006). The latter is unbiased, but has the problem that variables (such as the number of transitions between states) that are estimated from a posterior sample of stochastically mapped trees in which $\boldsymbol{Q}$ is set to its most likely value will tend to have variance that is slightly too low. Conversely, parameter estimates that we obtain from the full hierarchical Bayesian approach, in which $\boldsymbol{Q}$ and the stochastic histories are sampled from their joint posterior probability distribution, should generally have the correct variance; however, this approach depends on the somewhat difficult task of specifying a reasonable prior probability density for $\boldsymbol{Q}$.

When we've figured out the best approach for our tree and data, and then generated one or multiple stochastic maps, we can easily plot the stochastic character maps on a tree using different colors to map different character states through time. This is accomplished by computing the fraction of time spent in each state along each plotted edge in the tree. Having done this, we can then plot each state using a different colored line segment. For the lines connecting branches that share a common ancestor (i.e., the vertical or curved lines in Figs. 4.1a and b, respectively), we merely plot this line using the color of the last state on the preceding edge. (We could equally well use the initial state for any daughter edge, since under a continuous-time character evolution model, it is theoretically impossible that the character changes state exactly at a node.) Figure 4.3 shows an example circular tree for Greater Antillean anole species from Mahler et al. (2010) with a mapped discrete character "ecomorph"—the famous convergent ecological and morphological habitat specialists found in the *Anolis* lizard fauna of the Caribbean (Losos 2009). (The pie charts at internal nodes are not part of the stochastic map and will be

explained in Sect. 4.3.2, below.) To generate this stochastic map, I assumed a single substitution rate between all pairs of states, and then I fixed this rate to its most likely (i.e., maximum likelihood) value (Pagel 1994), making this an example of empirical Bayesian stochastic mapping (outlined above).

4.3.2 Aggregating Stochastic Maps: Node Posterior Probabilities

One of the difficulties that is inherent in visualizing the results of stochastic mapping using a plot like that of Fig. 4.3 is that plotting only one stochastic character map can create the misleading impression of certainty in the discrete character history. In fact, this plotted history is just one stochastic realization of many plausible histories, sampled in direct proportion to its Bayesian (or empirical Bayesian) posterior probability. Stochastic character mapping needs to be performed repeatedly (say, 100 or 1,000 times) to obtain a representative sample from the posterior distribution of plausible histories for our character; however, this creates the difficulty of having to somehow visualize in aggregate the results from many maps. If the character has changed state only once or a few times in the history of our tree, then the variability among stochastic reconstructions will generally be relatively small. In this case, any single stochastic map will be quite similar to the average map, and consequently, there may be no need to aggregate our visualization across maps. However, in cases when the character changes state more often, ancestral states at internal nodes and changes along branches will tend to vary much more among stochastically sampled reconstructions. In this case, it could be useful to employ a visualization technique that can incorporate information about this uncertainty.

One way to aggregate the results of many stochastic character histories is to simply compute the posterior probability that each node is in each state represented in our dataset. To do this, we just go through every branch of the tree and find the end state of that branch. The relative frequency of each character state for the ending state of each branch (except the root) is our estimate of the posterior probability that the corresponding daughter node is in that state. For the root node, we just pick one of the two daughter edges for each stochastic map simulation and compute the relative frequency that the starting state for that edge is in each state. It does not matter which one, because (again) it is theoretically impossible that the character changes state exactly at the root node in a continuous-time model.

To map these states on the tree, we can first plot our tree (or our tree with a representative stochastic character map) as normal. Then, having stored the horizontal and vertical positions of all internal nodes in memory, we can overlay the posterior probability that the node is in each state as a pie diagram plotted at each internal node of the tree using the same colors as were used in the mapping. The pie charts overlain on the graph of Fig. 4.3 show an example of this type of visualization.
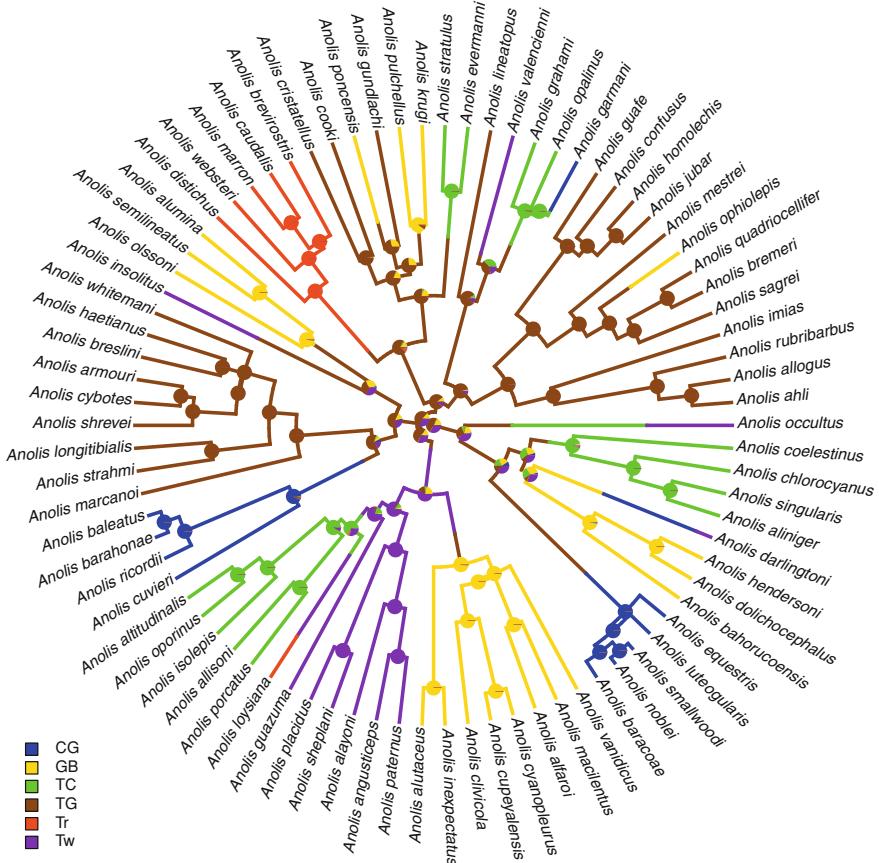
Fig. 4.3 Stochastic character mapping of the multistate discrete character "ecomorph" on the tree of Greater Antillean *Anolis* lizard ecomorph species. Any single map is a stochastic history sampled from the posterior distribution of histories in proportion to its probability. The pie charts at internal nodes show the posterior probabilities aggregated across 100 stochastically mapped character histories using the empirical Bayesian method and a single-rate (i.e., "equal rates") character evolution model. Ecomorphs are named for the microhabitat in which they are most often found, as follows: *CG* crown-giant, *GB* grass-bush, *TC* trunk-crown, *TG* trunk-ground, *Tr* trunk, *Tw* twig

4.3.3 Aggregating Stochastic Maps: Branch Posterior Density Mapping

If our discrete character is a binary (i.e., two-state) character, then we have a further option for plotting that does not completely dispose of the information in our stochastic maps about where along branches our character changes state. Instead, we can map the posterior probabilities of our binary discrete character

continuously along the branches and nodes of our phylogeny (Revell 2013). Figure 4.4a gives an illustration of the method using data for the feeding modes of biting vs. suction feeding in elopomorph eels (Collar et al. in revision).

To create this plot, I first finely segmented each branch of the maximum clade credibility tree from a Bayesian phylogeny inference posterior sample of 1,000 phylogenies. Then I went through all the stochastic character maps and asked whether each corresponding segment was in state *0* or state *1*. I tallied the relative frequency of each segment being entirely in state *0* or *1*, and for the instances in which the state changed within a segment, I computed a weighted tally with the weights being set equal to the fraction of time spent in each state for that segment. Having computed the posterior probability as a (near) continuous function of branch position, I then generated a color map and plotted the edges of the tree colored by this map. This method is implemented in phytools (Revell 2012, 2013).

We can use a similar approach to visualize the sampling variance across stochastic character maps in our posterior sample. For a binary character, this variance is equal to $p(1 - p)$, where p is the computed posterior probability that the character is in state *1* (vs. state *0*). If the character changes frequently on the tree, this will result in substantial variability in stochastic character histories. By contrast, if the character changes only once or a small number of times in the tree, then all stochastic maps will tend to be similar, and there will be much less variability among stochastic maps in the posterior sample. In Fig. 4.4b, I've created a plot showing the same phylogeny as in Fig. 4.4a with variability among stochastic reconstructions mapped along the branches in grayscale (from white being highly uncertain to black being highly certain). Note that the general pattern is that nodes and edges deep in the tree are uncertain, whereas nodes and edges close to the tips have low variance among maps. Although nodes and edges at the tips of the tree will always tend to have low variance, particularly as we get closer and closer to the tip states (which are known), as a general rule, deeper nodes and edges will be more uncertain if the character state changes frequently—and less so if it changes rarely.

Discerning readers may notice that no additional information is contained in Fig. 4.4b relative to Fig. 4.4a since any branch with a posterior probability of state *1* (biting) that is close to 0.0 or 1.0 will have low uncertainty (black), whereas any branch with intermediate posterior probability of being in state *1* will have high uncertainty (white). This is indeed correct. In fact, Figs. 4.4a and b are just different visualization of the same data, so it would be at the authors' discretion which is more appropriate to their study. It's interesting to consider that the conditions under which we can map the uncertainty of the posterior density from stochastic mapping on a tree are broader than the conditions in which we can map the posterior density itself. Specifically, it's technically challenging to map the posterior density for more than two (or perhaps three, see Revell 2013) stochastically mapped character states on the tree, whereas the difficulty of mapping the uncertainty among maps is not influenced by the number of states for our trait.
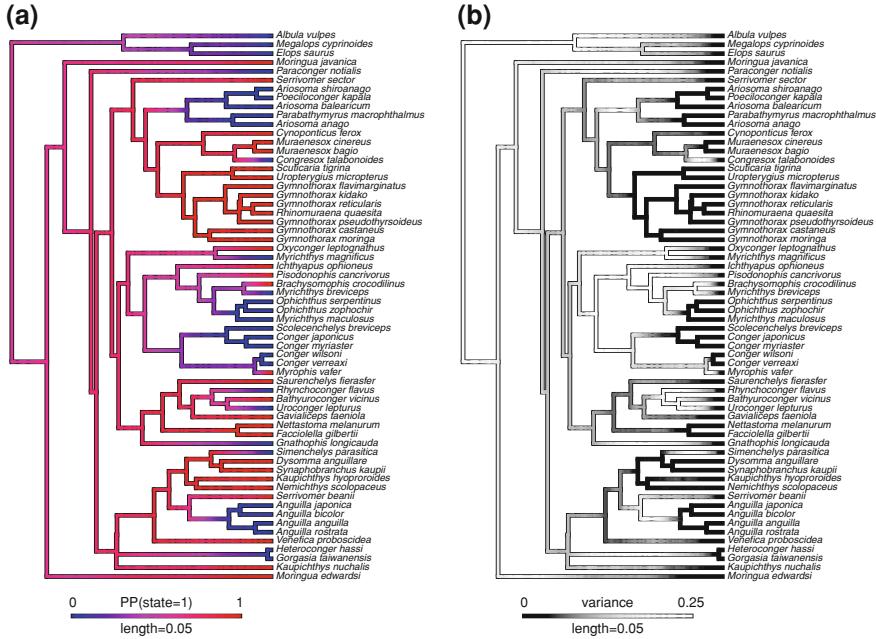
Fig. 4.4 Phylogeny of elopomorph eels from Collar et al. (in revision). **a** Posterior density of feeding mode "biting" versus "suction feeding" mapped on the tree from 100 stochastic character maps. **b** Variance among maps for feeding mode. Black indicates low variance (high certainty in the reconstructed trait value), whereas white indicates high variance

4.4 Continuous Character Methods

4.4.1 The Traitgram

One of the simplest approaches for visualizing continuous trait data on a tree is projection of the tree into a two-dimensional space defined by time and the continuous trait of interest. This has been called a "traitgram" (e.g., Ackerly 2009; Evans et al. 2009), and an example from simulated data is given in Fig. 4.5a.

The procedure to create a traitgram in this style is as follows. First, we estimate the ancestral states for our phenotypic trait at all the nodes of the tree (Schluter et al. 1997). Next, we compute for each node the height of the node above the root of the tree using a pre-order tree traversal. Then, we plot all nodes and tips with a vertical position determined by their known or estimated trait values and a horizontal position determined by the height above the root for that node or tip. Finally, we connect all parent and daughter nodes by edges. In this type of visualization, it's important to keep in mind only the horizontal dimension of edge length contains information about branching times in the tree.
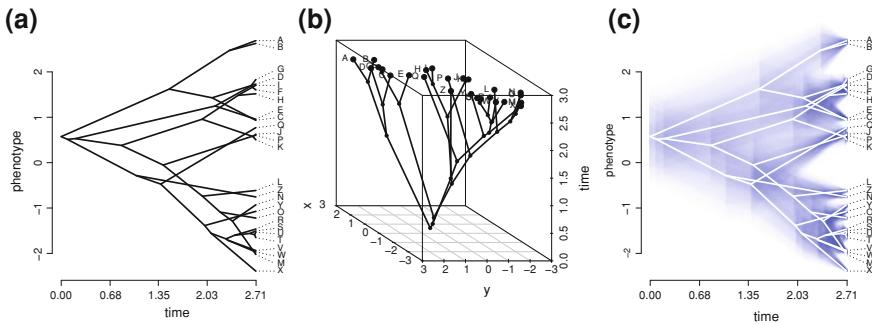
Fig. 4.5 a Hypothetical "traitgram" (projection of the tree into a space defined by time and the continuous trait) for a simulated, 26-taxon tree. **b** Simulated three-dimensional traitgram (two phenotypic trait axes plus time). **c** The traitgram from Fig. 4.5a, with 95 % confidence limits around ancestral values shown by increasing transparency in the plotted lines

Usually, we label terminal nodes using the names of the corresponding tip species in the tree. If a number of tips have similar values on the phenotypic trait axis, then this can create a messy plot in which labels overlap and are unreadable. In the example of Fig. 4.5a, I used numerical optimization to space the tip labels using a cost function that penalizes both label overlap and distance from the vertical position of the tip node. (We can specify arbitrary costs for each—depending on whether we find label overlap or label vertical displacement more undesirable.)

It is also possible create a three-dimensional traitgram. In this case, the vertical dimension (for example) might be time, whereas the remaining two dimensions show observed or reconstructed trait values for two continuously valued traits. Figure 4.5b shows a static image of a three-dimensional traitgram plotted with time on the vertical axis. Although this is a fixed plot, phytools can also create a three-dimensional plotting object that can be spun or animated using the R package rgl (Adler and Murdoch 2013).

Finally, the traitgram algorithm can be used to create a visualization capturing uncertainty in ancestral character estimation. In this case, we can use the Hessian matrix or the formulae of Rohlf (2001) to compute the standard error and 95 % confidence interval around ancestral state estimates. Having done this, we can plot 95 % confidence limits around ancestral states (and edges in the traitgram) using a continuous color or transparency gradient, such as that illustrated in Fig. 4.5c.

4.4.2 Projection of a Tree Into Morphospace

Another common visualization method is a complete projection of the tree into a two- or three-dimensional morphospace. This visualization is commonly referred to as a "phylomorphospace" plot (e.g., Rohlf 2002; Sidlauskas 2008).
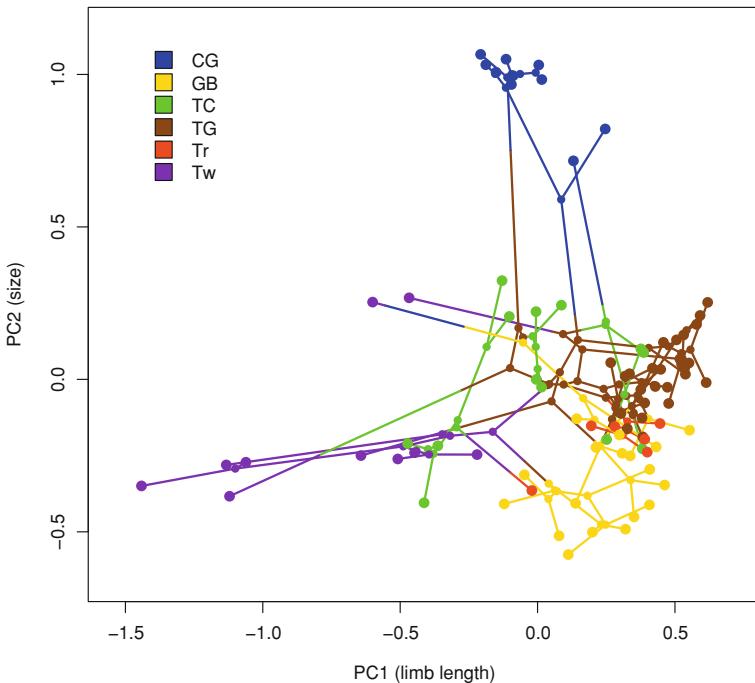
Fig. 4.6 Phylomorphospace (projection of the tree into morphospace) for two principal component axes from 82 species of Greater Antillean *Anolis* lizards. "Ecomorph" class (i.e., ecomorphological habitat specialist) from one stochastic map is projected onto the phylomorphospace. Ecomorphs are as in Fig. 4.3

To create a phylomorphospace visualization, we first need to estimate the ancestral states for all internal nodes in the tree including the root (Schluter et al. 1997). Then, we plot all the estimated states at nodes and the observed states at the tips into our bivariate space. Finally, we connect all parent to daughter nodes with edges and add labels if desired. A visualization of a phylomorphospace plot in two dimensions is given in Fig. 4.6. This figure shows a projection of the phylogeny of greater Antillean ecomorph species of anoles into a two-dimensional principal component morphospace defined by relative limb lengths on the horizontal (PC1) and overall size on the vertical (PC2). Overlain on the projection is a single stochastic map of the evolution of ecomorph on the tree of anoles—the same stochastic map, in fact, as in Fig. 4.3. (See Sect. 4.3.1 for more detail on stochastic character mapping.) The phylogeny and data are from Mahler et al. (2010).

One unfortunate attribute of phylomorphospace visualizations of this type is that all information about time since the root is thrown away during plotting. Recently, some authors (Miller et al. 2013) developed an approach to try and show this information on a phylomorphospace plot using a color gradient that changes continuously from the root to the tips of the tree. Figure 4.7 shows a tree (in panel *a*)
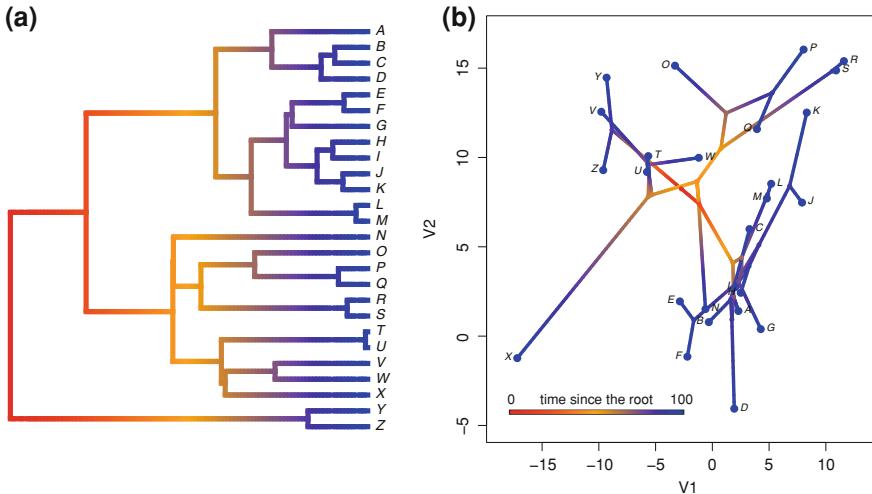
Fig. 4.7 a Stochastic phylogeny with time since the root overlain as a color gradient. **b** Simulated phylomorphospace with the color gradient retained (following Miller et al. 2013)

and an example phylomorphospace plotted in this style (Fig. 4.7b), with the temporal dimension retained via a continuous color gradient from the root of the tree (red) toward the tips (blue).

4.4.3 Continuous Character Mapping on the Tree

A final method for continuous character visualization uses the same technique as was described for a posterior density plot from a set of stochastic map trees; however, in this case, we estimate the ancestral states for internal nodes using ML and then interpolate the states along the branches of the tree using Eq. (2) from Felsenstein (1985). Having done this, we are prepared to map our continuous trait on the tree using a continuous color gradient. This method is implemented in phytools (Revell 2012, 2013; also see Verbruggen 2008). An example of this continuous trait mapping is given in Fig. 4.8 using log-transformed body size (snout-to-vent length, or SVL) in Greater Antillean *Anolis* lizards (Mahler et al. 2010). Using the phytools package, it is also possible to combine methods a and b of this section to visualize trait evolution for more than two characters in a single graph. For instance, Fig. 4.9 shows a four-trait multivariate phylogenetic scatterplot matrix for simulated data. The diagonal consists of continuous character maps for each *i*th trait, whereas the *i, j*th off-diagonal cell shows a bivariate projection of the tree into morphospace for traits *i* and *j*. For both Figs. 4.8 and 4.9, the specific color scheme is arbitrary, and a different color palette can easily be specified by the user (if, for instance, a more color-blind-sensitive color scheme is desired).
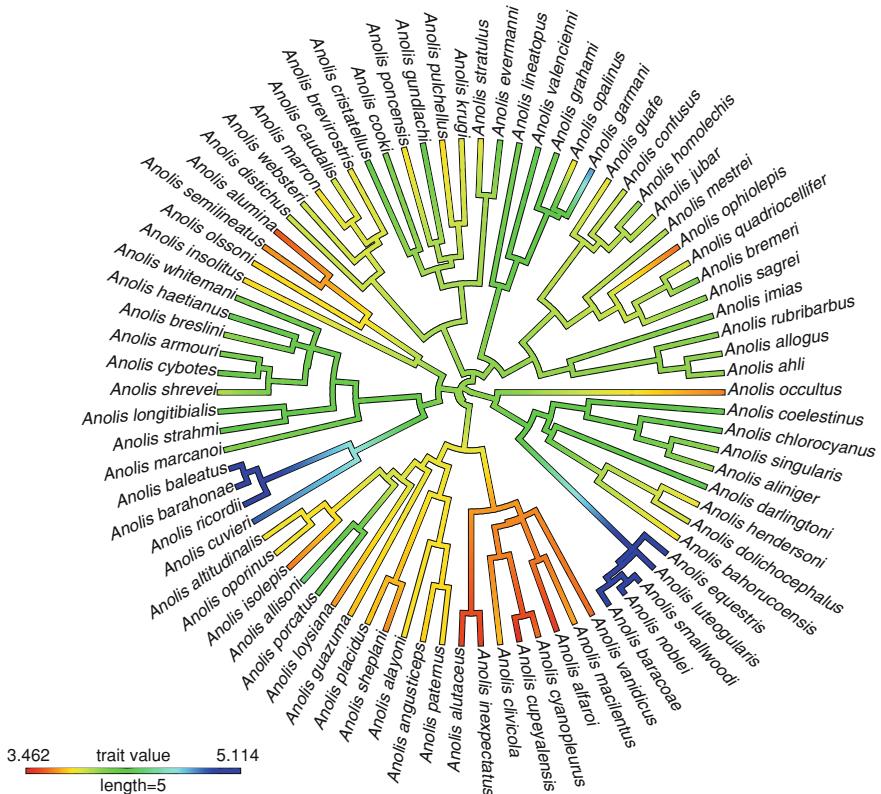
Fig. 4.8 Body size mapped using a continuous color gradient on a phylogeny of 82 Caribbean anole species from Mahler et al. (2010)

4.5 Additional Methods

In Sects. 4.3 and 4.4, I illustrated some different visualization techniques for discrete and continuous character data; however, it is also relatively straightforward to combine some of these techniques into a single plot. I already showed an example of this in Fig. 4.6, which gives a phylomorphospace with an overlain discrete character stochastic mapping.

  As these approaches thus far only extend in minor ways the visualizations already shown, I won't dwell extensively on specific methodology; however, by way of illustration, Fig. 4.10a shows a stochastically mapped discrete character overlain on a continuous character traitgram, while Fig. 4.10b shows a bivariate projection of the tree into morphospace with a posterior density from stochastic mapping overlain. This type of plotting method is especially useful in exploratory data analysis for datasets in which (for instance) the state of discrete character is hypothesized to influence the rate of evolution in a continuous character (e.g., in
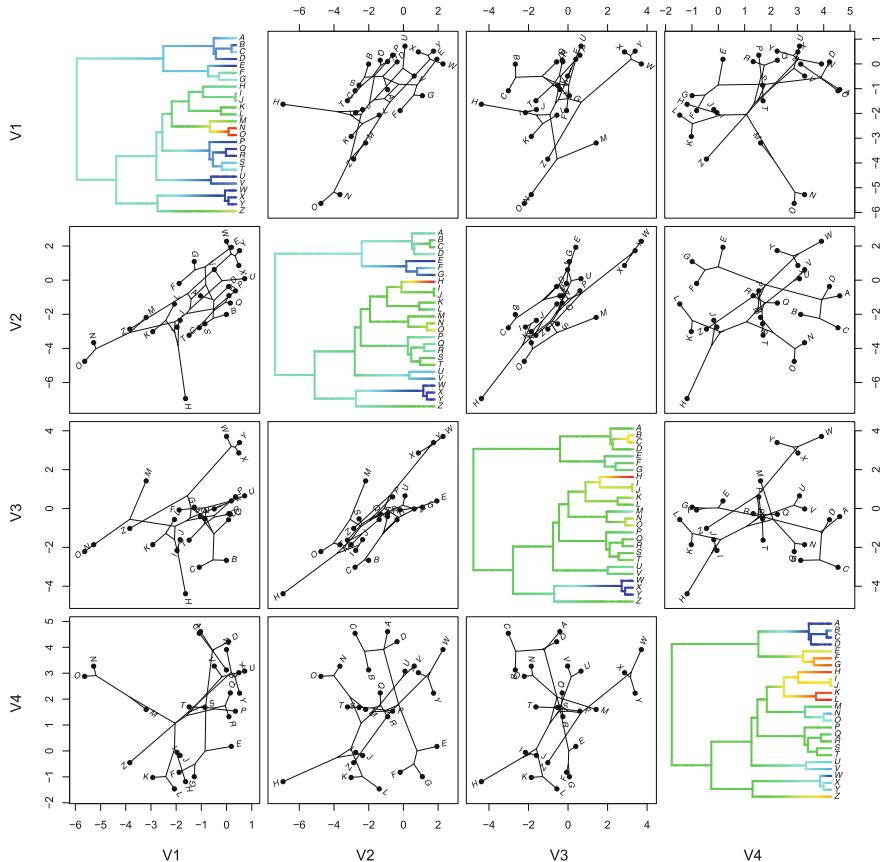
Fig. 4.9 Simulated four-trait phylogenetic scatterplot matrix. Each diagonal element is a continuous character projection on the tree in which red branches indicate small values for the phenotypic trait and blue branches large values, whereas off-diagonals are phylomorphospaces for each i, jth pair of traits

the simulated data of Fig. 4.10a), or in which the state of a discrete character influences the evolutionary correlation between traits (e.g., Fig. 4.10b).

Finally, trees can also be projected onto a geographic map. For instance, Fig. 4.11a shows a simulated phylogenetic tree in which the tips of the tree point to different geographic localities (perhaps the center of a hypothetical species range or the type locality for the species) on a world map. All the nodes on the tree have been rotated using a "greedy" optimization method to minimize line crossing. The method merely climbs up the tree using a pre-order (root-to-tip) traversal, rotates each node, and accepts the rotation if it reduces the objective function—which is the difference in rank order between the left-to-right order of the tip labels and the west-to-east ordering on the map. Figure 4.12b shows a different type of direct projection of the tree onto a map; however, in this case, it is
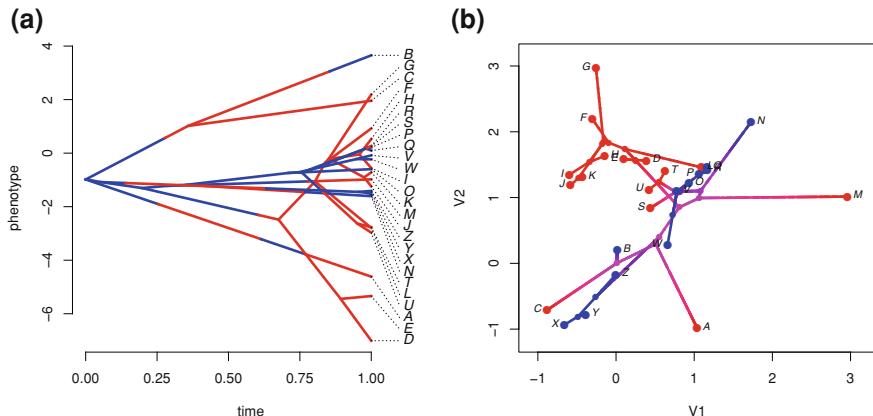
Fig. 4.10 a Traitgram on a simulated tree with a stochastic character map overlain. The continuous character data were simulated with a high rate of evolution on the red branches of the tree and a low rate on the blue branches. **b** Phylomorphospace with a posterior density map from 100 stochastic character maps overlain. Data were simulated with a high evolutionary correlation on the blue branches of the tree

important to keep in mind that the locations of the internal nodes in this projection are not equivalent to ancestral range reconstructions (which may be possible to obtain using different methods outside the scope of this chapter, e.g., Ree and Smith 2008).

4.6 Programming Phylogeny Visualization Methods in R

4.6.1 The Structure of a "phylo" Object

The first and most useful thing to understand when developing a plotting method for phylogenies in R is the basic structure of a phylogeny in memory. Phylogenies are stored in R as an object of type *list* with the *class* attribute set to "phylo". Thus, we say that a phylogeny is stored as an *object of class "phylo"*. A list in R consists of a set of objects that can be the same or different in type. For instance, a list could consist of a matrix, a vector of real numbers, and a character string. In this case, the "phylo" object, *tree*, consists of the following four elements and one or more attributes. We can denote the elements of a list using double square brackets (i.e., [[]]) or the dollar sign ($):

tree$edge: *edge* is a matrix of dimensions $e \times 2$ in which e is the number of edges in the tree. For a fully bifurcating tree, $e = 2n - 2$, for n total taxa in the tree. This is just the number of tips (n) plus the number of internal nodes ($n - 1$) minus 1, since every tip or internal node (except for the root) is preceded by an edge. The matrix
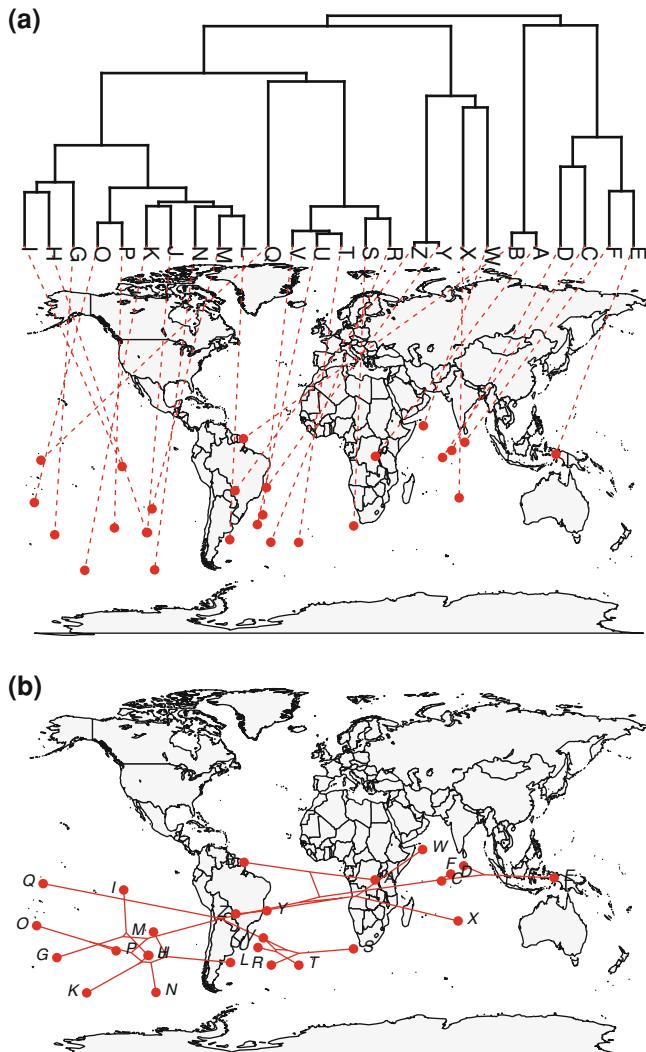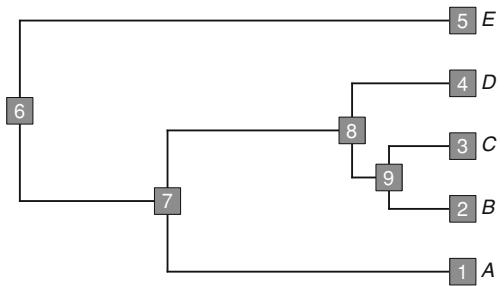
Fig. 4.11 Two different projections of a phylogeny onto a geographic map. **a** A projection in which the tips of the tree are connected to locations on the map via dotted lines. The nodes of the tree were rotated using a "greedy" optimization method to minimize line crossing. **b** A "direct" projection of the phylogeny onto the tree. Note that nodes should not be interpreted as reconstructed ancestral areas in this visualization

tree$edge contains the starting and ending indices (i.e., node number) of every edge in the tree. These indices are given in the boxed numbers of the example five-taxon tree of Fig. 4.12. By convention, indices 1 through n are assigned to the tip nodes in the tree, whereas indices $n + 1$ through $m + n + 1$ (for m internal nodes) are

Fig. 4.12 Five-taxon
phylogeny with node
numbers

assigned to the internal nodes of the tree. For the tree in Fig. 4.12, an example
ordering of *tree$edge* is as follows:

```
> tree$edge
     [,1] [,2]
[1,]    6    7
[2,]    7    1
[3,]    7    8
[4,]    8    9
[5,]    9    2
[6,]    9    3
[7,]    8    4
[8,]    6    5
```

tree$Nnode: *Nnode* is an integer giving the total number of internal nodes in the
tree, including the root.

```
> tree$Nnode
[1] 4
```

tree$tip.label: *tip.label* is a vector containing all the tip labels for the tips of the
tree. The order of *tree$tip.label* is the index order for the nodes. For instance, for
the tree in Fig. 4.12, this is merely as follows:

```
> tree$tip.label
[1] "A" "B" "C" "D" "E"
```

tree$edge.length: *edge.length* is a vector containing the lengths of all the edges of
the tree in the order of the rows of *tree$edge*. In Fig. 4.12, this vector is as follows:

```
> tree$edge.length
[1]  4  8  5  1  2  2  3 12
```

Finally, the "phylo" object has at least one attribute, its class. This is just a
string which tells R how to treat the object in certain custom functions built to deal
with objects of this type. In this case, the attribute is simply:

```
> attr(tree,"class")
[1] "phylo"
```

Special types of "phylo" objects can have additional elements or attributes.

4.6.2 Plotting a Simple Phylogram

The next thing that I'll illustrate is how to use the algorithm of Sect. 4.2 to plot a simple, right-facing phylogram. Whereas in Sect. 4.2 I focused on a general algorithm that applies theoretically to any programming language or development environment, here I'll give specific R code. Obviously, packages like the R phylogenetics libraries ape and phytools already contain numerous functions for drawing trees; however, a basic understanding of how trees are plotted in R may be useful to investigators interested in developing totally new approaches for visualization.

The code I give below *depends* on the R packages ape and phytools. That means that it uses functions internally that belong to those R function libraries. To start, we should load those packages:

```
library(ape)
library(phytools)
```

The first step is to figure out how many tips we have in the tree and then reorder the tree so that the edges of *tree$edge* are "cladewise"—that is, edges in the same clade are next to each other in the matrix:

```
n <- length(tree$tip.label)
cw <- reorder(tree,"cladewise")
```

Next, we want to compute the vertical position of all the edges in our rightward-facing tree. To do this, we assign our cladewise-ordered tip heights 1 through *n* and then compute the heights for all internal edges via one post-order tree traversal:

```
## create vector
y <- vector(length=n+cw$Nnode)
## assign heights for tips
y[cw$edge[cw$edge[,2]<=n,2]] <- 1:n
pw <- reorder(tree,"pruningwise")
nn <- unique(pw$edge[,1])
## assign heights for internal nodes
for(i in 1:length(nn)){
  yy <- y[pw$edge[which(pw$edge[,1]==nn[i]),2]]
  y[nn[i]] <- mean(range(yy))
}
```

Then, we compute the starting and ending points of each edge on the tree. This can be done for a tree in cladewise order using the phytools function *nodeHeights*:

```
X <- nodeHeights(cw)
```

The matrix *X* has dimensions equal to *tree$edge,* and every element of *X* corresponds to the height above the root of the corresponding element of *tree$edge*.

Now, we are ready to open a new plotting object. Here, we crudely size the horizontal (*x*) dimension of our plotting area to be 10 % larger than the total length of our tree—to allow space for labels. In "real" tree-plotting functions, we would use a more sophisticated algorithm for this to ensure that enough (but not too much) space was allocated for plotting labels:

```
plot.new()
par(mar=rep(0.1,4))
plot.window(xlim=c(0,1.1*max(X)),ylim=c(0,max(y)+1))
```

Next, we can plot all the horizontal lines in our tree. This is easy because for each edge, the *x*-axis coordinates correspond to a row of *X*. The single *y*-coordinate can be found by matching the endpoint of the edge (i.e., *cw$edge[i,2]* for the *i*th edge) with the vector *y*:

```
for(i in 1:nrow(X)) lines(X[i,],rep(y[cw$edge[i,2]],2),lwd=2,lend=2)
```

Then, we add the vertical lines that show the relationships between taxa. Only internal nodes have vertical lines, so we just go through the indices used for internal nodes. Each time, we find the element of *X* and the coordinates from *y* that correspond with the target edge, and plot the following:

```
for(i in 1:tree$Nnode+n)
  lines(X[which(cw$edge[,1]==i),1],range(y[cw$edge[which(cw$edge[,1]==i),2]]),lwd=2,
  lend=2)
```

Finally, we can plot tip labels. This is easy. The vertical position is the position we assigned at the beginning of the exercise; the horizontal position is the corresponding node height of the terminal node for that tip:

```
for(i in 1:n) text(X[which(cw$edge[,2]==i),2],y[i],tree$tip.label[i],pos=4,offset=0.1)
```

Try it!

4.6.3 Plotting a Simple Projection of the Tree Into Morphospace (Phylomorphospace)

In Sect. 4.4.2, I described a method to project a phylogenetic tree into a two-dimensional morphospace (i.e., a phylomorphospace plot). What follows is a bit more detail on how to program this visualization method in R, which, as the reader

will see, is even simpler than plotting a phylogram. This function again uses ape and phytools. Since phytools is dependent on ape, simply loading phytools should be sufficient:

```
library(phytools)
```

First, let's calculate how many tips we have and then estimate ancestral states for all internal nodes. The latter is accomplished using the phytools function *fastAnc*:

```
n <- length(tree$tip.label)
x <- c(x[tree$tip.label],fastAnc(tree,x))
y <- c(y[tree$tip.label],fastAnc(tree,y))
```

Now, let us plot the tips and nodes of our tree. For better ease of visualization, let's plot internal nodes with a slightly smaller symbol than tips:

```
plot(x[1:n],y[1:n],cex=1.25,pch=21,bg="black",xlab="x",ylab="y")
points(x[1:tree$Nnode+n],y[1:tree$Nnode+n],cex=1,pch=21,bg="black")
```

Then, add the lines connecting parent and daughter nodes in morphospace.

```
apply(tree$edge,1,function(edge,x,y) lines(x[edge],y[edge]),x=x,y=y)
```

Finally, let's label all terminal nodes:

```
text(x[1:n],y[1:n],tree$tip.label,pos=2)
```

4.7 Conclusions and Future Directions

Phylogenetic comparative methods have become central to evolutionary biology over the past thirty or so years (Miles and Dunham 1993; Freckleton et al. 2002; Losos 2011; Baum and Smith 2013) and have even begun to infiltrate other biological and non-biological disciplines, such as genomics, biological anthropology, and linguistics (e.g., Thornton and Desalle 2000; Atkinson and Gray 2005; Nunn 2011). Many chapters of this book discuss innovative new approaches for data analysis in comparative biology. However, an important—but sometimes over-looked—first and last step in data analysis is often visualization. *First*, because plotting trees and comparative data can alert us to deviations or errors in our data and perhaps suggest methods of study that might be useful for our data or question. For instance, in a continuous character mapping on the tree, a color gradient along an edge of the tree that suggested that a lineage changed from the highest observed value of the trait to the lowest (or vice versa) might inspire us to cross-check the

phenotypic trait value in our dataset, or the position of a potentially "rogue" lineage in the tree. *Last*, because presenting persuasive and informative figures can be an important tool in conveying relevant information about our study system, question, and results.

Potential methods for visualizing phylogenies and comparative data are limited only by the scope of our imaginations (e.g., Rosindell and Harmon 2012). In this article, I have concentrated on relatively simple methods implemented in one way or another in the phytools R package (Revell 2012). Some of these were originally devised by me, but others were devised by others and implemented by me (e.g., Rohlf 2002; Sidlauskas 2008; Evans et al. 2009; Miller et al. 2013). Other methods still were devised in a slightly different form by others and adapted by me for R and the phytools package (e.g., Verbruggen 2008). The list of methods described in this chapter is not comprehensive; however, it does sample from a broad swath of approaches for visualization in phylogenetic comparative biology across discrete and continuous character data types. I have not discussed visualization methods that use the tree but no phenotypic trait data for comparative analysis (for instance, lineage-through-time plots; Pybus and Harvey 2000; Harmon et al. 2003). A review of these methods could be the topic of a separate article or book chapter.

One major limitation of the approaches described in this chapter is that they are constrained to the "paper paradigm" (Rosindell and Harmon 2012). That is, they are designed to be printed on a piece of paper. The printed page (or at least an electronic version thereof) continues to be the primary mode of communication in the sciences. However, this medium imposes severe limits on the size and scope of visualizations of comparative data and phylogenies. Phylogenetic datasets can now contain thousands or perhaps even tens of thousands of tips (e.g., Bininda-Emonds et al. 2007; Smith et al. 2009). Most of the methods of this chapter would be ineffective at conveying useful information about phylogenetic comparative data for phylogenies of this size. Future method development in phylogenetic comparative biology should look to move beyond the paper paradigm for solutions in visualizing large phylogenies and multivariable phenotypic datasets.

Phylogenetic comparative biology has grown over the past thirty or so years to assume a central position in evolutionary study (Miles and Dunham 1993; Losos 2011). Along with it have come new challenges in visualizing comparative data on trees. In this chapter, I have discussed a number of novel or newly implemented visualization methods for comparative data and phylogenies. In the future, new approaches must address the challenge of very large phylogenies (e.g., Rosindell and Harmon 2012) and increasingly multivariate phenotypic trait data of modern phylogenetic studies.

References

Ackerly D (2009) Conservatism and diversification of plant functional traits: evolutionary rates versus phylogenetic signal. PNAS 106:19699–19706

Adler D, Murdoch D (2013) rgl: 3D visualization device system (OpenGL). R package version 0.93.935

Archie J, Day WHE, Felsenstein J, Maddison W, Meacham C, Rohlf FJ, Swofford D (1986) Newick tree format. More information: http://evolution.genetics.washington.edu/phylip/newicktree.html

Atkinson QD, Gray RD (2005) Curious parallels and curious connections: phylogenetic thinking in biology and historical linguistics. Syst Biol 54:513–526

Baum DA, Smith SD (2013) Tree thinking: an introduction to phylogenetic biology. Roberts and Company, Greenwood Village

Beaulieu JM, O'Meara BC, Donoghue MJ (2013) Identifying hidden rate changes in the evolution of a binary morphological character: the evolution of plant habit in the campanulid angiosperms. Syst Biol 62:725–737

Becker RA, Wilks AR, Brownrigg R, Minka TP (2013) Maps: draw geographical maps. R package version 2.3-2

Bininda-Emonds OR, Cardillo M, Jones KE, MacPhee RDE, Beck RMD, Grenyer R, Price SA, Vos RA, Gittleman JL, Purvis A (2007) The delayed rise of present-day mammals. Nature 446:508–512

Bokma F (2008) Detection of "punctuated equilibrium" by Bayesian estimation of speciation and extinction rates, ancestral character states, and rates of anagenetic and cladogenetic evolution on a molecular phylogeny. Evolution 62:2718–2726

Bollback JP (2006) SIMMAP: stochastic character mapping of discrete traits on phylogenies. BMC Bioinf 7:88

Butler MA, King AA (2004) Phylogenetic comparative analysis: a modeling approach for adaptive evolution. Am Nat 164:683–695

Collar DC, Wainwright PC, Alfaro ME, Revell LJ, Mehta RS, Biting disrupts integration to spur skull evolution in eels

Core Team R (2013) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Eastman JM, Alfaro ME, Joyce P, Hipp AL, Harmon LJ (2011) A novel comparative method for modeling shifts in the rate of character evolution on trees. Evolution 65:3578–3589

Edwards EJ, Osborne CP, Strömberg CAE, Smith SA, C4 Grasses Consortium (2010) The origins of C4 grasslands: Integrating evolutionary and ecosystem science. Science 328:587–591

Evans MEK, Smith SA, Flynn RS, Donoghue MJ (2009) Climate, niche evolution, and diversification of the "bird-cage" evening primroses (Oenothera, sections Anogra and Kleinia). Am Nat 173:225–240

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

Felsenstein J (1988) Phylogenies and quantitative characters. Ann Rev Ecol Syst 19:445–471

Felsenstein J (2004) Inferring phylogenies. Sinauer Associates, Sunderland

Felsenstein J (2012) A comparative method for both discrete and continuous characters using the threshold model. Am Nat 179:145–156

FitzJohn RG (2010) Quantitative traits and diversification. Syst Biol 59:619–633

Freckleton RP, Harvey PH, Pagel M (2002) Phylogenetic analysis and comparative data: a test and review of evidence. Am Nat 160:712–726

Glor RE (2010) Phylogenetic insights on adaptive radiation. Ann Rev Ecol Evol Syst 41:251–270

Harmon LJ, Schulte JA II, Larson A, Losos JB (2003) Tempo and mode of evolutionary radiation in iguanian lizards. Science 301:961–964

Harvey PH, Pagel MD (1991) The comparative method in evolutionary biology. Oxford University Press, Oxford

Huelsenbeck JP, Nielsen R, Bollback JP (2003) Stochastic mapping of morphological characters. Syst Biol 52:131–158

Lemon J (2006) Plotrix: A package in the red light district of R. R-News 6:8–12

Ligges U, Mächler M (2003) Scatterplot3d: An R Package for visualizing multivariate data. J Stat Softw 8:1–20

Losos JB (2009) Lizards in an evolutionary tree: ecology and adaptive radiation of anoles. University of California Press, Berkeley

Losos JB (2011) Seeing the forest for the trees: the limitations of phylogenies in comparative biology. Am Nat 177:709–727

Mahler DL, Revell LJ, Glor RE, Losos JB (2010) Ecological opportunity and the rate of morphological evolution in the diversification of Greater Antillean anoles. Evolution 64:2731–2745

Miles DB, Dunham AE (1993) Historical perspectives in ecology and evolutionary biology: the use of phylogenetic comparative analyses. Ann Rev Ecol Syst 24:587–619

Miller ET, Zanne AE, Ricklefs RE (2013) Niche conservatism constrains australian honeyeater assemblages in stressful environments. Ecol Lett 16:1186–1194

Near TJ, Bolnick DI, Wainwright PC (2005) Fossil calibrations and molecular divergence time estimates in centrarchid fishes (Teleostei: Centrarchidae). Evolution 59:1768–1782

Nielsen R (2002) Mapping mutations on phylogenies. Syst Biol 51:729–739

Nunn CL (2011) The comparative approach in evolutionary anthropology and biology. University of Chicago Press, Chicago

O'Meara B (2012) Evolutionary inferences from phylogenies: a review of methods. Ann Rev Ecol Evol Syst 43:267–285

O'Meara BC, Ané C, Sanderson MJ, Wainwright PC (2006) Testing for different rates of continuous trait evolution using likelihood. Evolution 60:922–933

Pagel M (1994) Detecting correlated evolution on phylogenies: a general method for comparative analysis of discrete characters. Proc Roy Soc Ser B 255:37–45

Paradis E (2012) Analysis of phylogenetics and Evolution with R, 2nd edn. Springer, New York

Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290

Pennell MW, Harmon LJ (2013) An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology and paleobiology. Ann NY Acad Sci 1289:90–105

Pybus OG, Harvey PH (2000) Testing macro-evolutionary models using incomplete molecular phylogenies. Proc Roy Soc B 267:2267–2272

Ree RH, Smith SA (2008) Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. Syst Biol 57:4–14

Revell LJ (2012) Phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol Evol 3:217–223

Revell LJ (2013) Two new graphical methods for mapping trait evolution on phylogenies. Methods Ecol Evol 4:754–759

Revell LJ (2014) Ancestral character estimation under the threshold model from quantitative genetics. Evolution 68:743–759

Revell LJ, Mahler DL, Peres-Neto PR, Redelings BD (2012) A new method for identifying exceptional phenotypic diversification. Evolution 66:135–146

Rohlf FJ (2001) Comparative methods for the analysis of continuous variables: geometric interpretations. Evolution 55:2143–2160

Rohlf FJ (2002) Geometric morphometrics and phylogeny. In: MacLeod N, Forey PL (eds) Morphology, shape, and phylogeny. CRC Press: Boca Raton, pp 175–193

Rosindell J, Harmon LJ (2012) OneZoom: a fractal explorer for the tree of life. PLoS Biol 10:e1001406

Schliep KP (2011) Phangorn: phylogenetic analysis in R. Bioinformatics 27:592–593

Schluter D, Price T, Mooers AØ, Ludwig D (1997) Likelihood of ancestor states in adaptive radiation. Evolution 51:1699–1711

Sidlauskas B (2008) Continuous and arrested morphological diversification in sister clades of characiform fishes: a phylomorphospace approach. Evolution 12:3135–3156

Smith SA, Beaulieu JM, Donoghue MJ (2009) Mega-phylogeny approach for comparative biology: an alternative to supertree and supermatrix approaches. BMC Evol Biol 9:37

Thornton JW, DeSalle R (2000) Gene family evolution and homology: genomics meets phylogenetics. Ann Rev Genomics Hum Genet 1:41–73

Verbruggen H (2008) TreeGradients. Available: http://www.phycoweb.net/software/TreeGradients/

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

Chapter 5
A Primer on Phylogenetic Generalised Least Squares

Matthew R. E. Symonds and Simon P. Blomberg

Abstract Phylogenetic generalised least squares (PGLS) is one of the most commonly employed phylogenetic comparative methods. The technique, a modification of generalised least squares, uses knowledge of phylogenetic relationships to produce an estimate of expected covariance in cross-species data. Closely related species are assumed to have more similar traits because of their shared ancestry and hence produce more similar residuals from the least squares regression line. By taking into account the expected covariance structure of these residuals, modified slope and intercept estimates are generated that can account for interspecific autocorrelation due to phylogeny. Here, we provide a basic conceptual background to PGLS, for those unfamiliar with the approach. We describe the requirements for a PGLS analysis and highlight the packages that can be used to implement the method. We show how phylogeny is used to calculate the expected covariance structure in the data and how this is applied to the generalised least squares regression equation. We demonstrate how PGLS can incorporate information about phylogenetic signal, the extent to which closely related species truly are similar, and how it controls for this signal appropriately, thereby negating concerns about unnecessarily 'correcting' for phylogeny. In addition to discussing the appropriate way to present the results of PGLS analyses, we highlight some common misconceptions about the approach and commonly encountered problems with the method. These include misunderstandings about what phylogenetic signal refers to in the context of PGLS (residuals errors, not the traits themselves), and issues associated with unknown or uncertain phylogeny.

M. R. E. Symonds (✉)
Centre for Integrative Ecology, School of Life and Environmental Sciences,
Deakin University, Burwood, VIC, Australia
e-mail: matthew.symonds@deakin.edu.au

S. P. Blomberg
School of Biological Sciences, The University of Queensland, St Lucia, QLD, Australia

5.1 Introduction

5.1.1 The Background to PGLS

The 1980s saw a rise in appreciation of the need to take phylogeny into account when conducting analyses of trait correlations across species (Ridley 1983; Felsenstein 1985; Huey 1987; Harvey and Pagel 1991; for an entertaining overview see Losos 2011). Because of shared evolutionary history, species do not provide independent data points for analysis, thereby violating one of the fundamental assumptions of most statistical tests (Chap. 1). With appreciation of this problem came the impetus to develop statistical methods for analysing comparative data while taking phylogeny into account. Of these, phylogenetic generalised least squares (PGLS) is one of the primary methods employed.

PGLS (also called 'phylogenetic regression' or 'phylogenetic general linear models') was a method initially formulated by Grafen (1989) and subsequently developed by Martins and Hansen (1997), Pagel (1997, 1999) and Rohlf (2001). Initially, biologists were slow to incorporate phylogenetic comparative methods in their research, perhaps because methodological papers plunge quickly into mathematical formulae and statistical terminology. This chapter is intended for those without a strong statistical background as an introduction to PGLS. We explain how PGLS incorporates information about phylogeny and the strength of the phylogenetic signal: the extent to which closely related species resemble each other. We will provide advice on how to conduct analyses, and present results, and also point out areas where those new to the methods might get stuck.

5.1.2 What Kind of Analyses are PGLS Used for?

The most common type of analyses where PGLS are employed are those which seek to establish the nature of the evolutionary association between two or more biological traits—for example, the relationship between body mass and life span (Promislow and Harvey 1990). By 'evolutionary association', we mean evidence that traits are associated over evolutionary time. Although PGLS is frequently used to examine the association between a pair of traits, it can also handle multiple predictor variables. However, PGLS has a wider range of applications, including ancestral state estimation, assessment of mode of evolution, and identification of directionality of evolution among traits.

Analyses of coevolution among traits typically involve the estimation of regression estimates. For PGLS, the dependent (response) variable is usually a continuous variable. The predictor variable(s) may also be continuous, but PGLS can deal with pseudo-continuous ordinal data and binary discrete data. Multi-state discrete variables with non-ordinal properties (e.g. diet: insectivorous, herbivorous, piscivorous, etc.) can be dealt within a PGLS framework if they are recoded as separate binary characters (e.g. piscivory: no (0) or yes (1)).
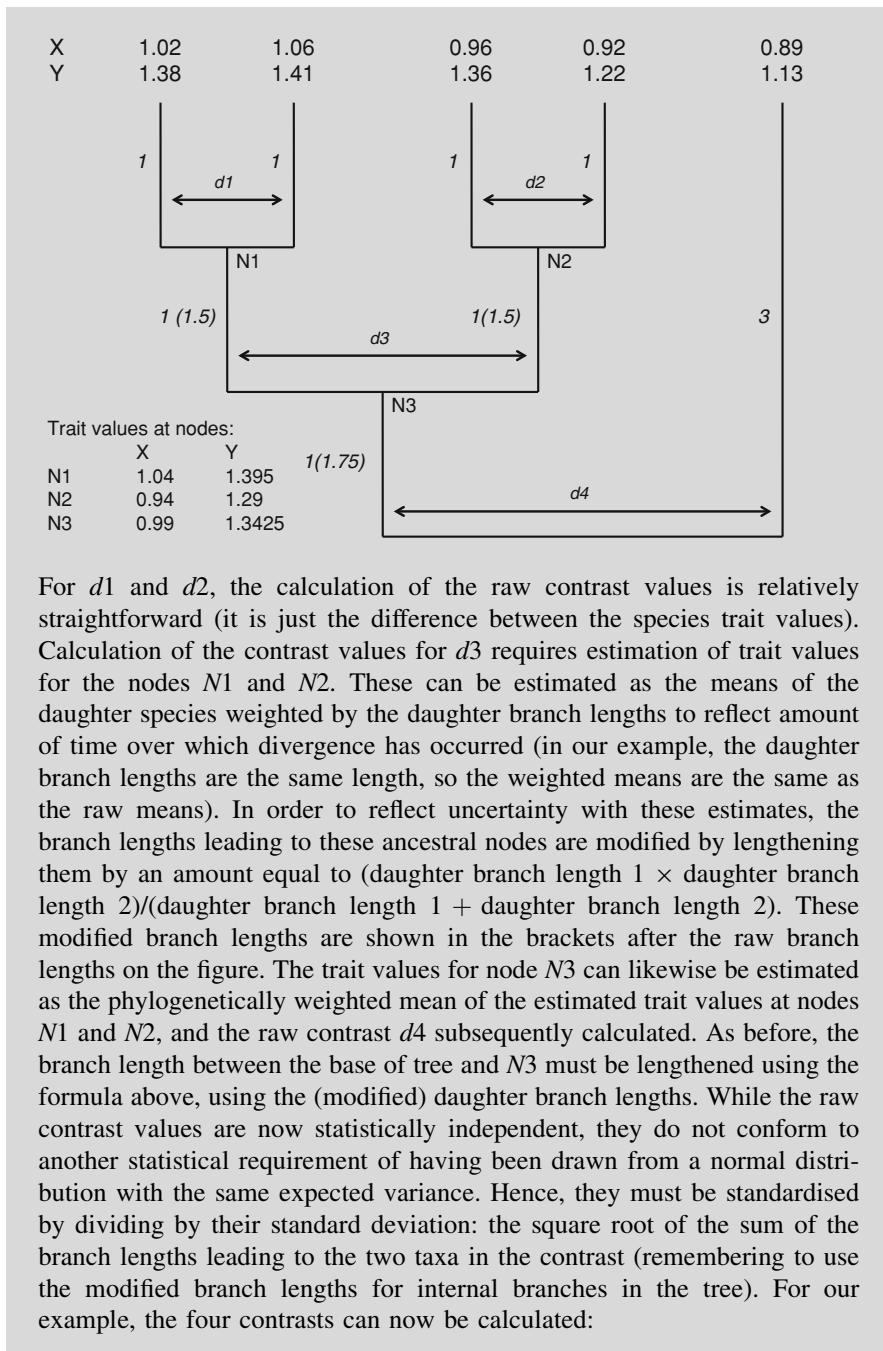
Hypothesis testing with PGLS is not appropriate for analyses with a discrete character as the response variable. Separate methods exist for dealing with discrete response variables including the concentrated changes test (Maddison 1990), pairwise comparisons (Maddison 2000), Pagel's (1994) likelihood method, and phylogenetic logistic regression (Ives and Garland 2010). Chapter 9 reviews some of these approaches.

5.1.3 PGLS and Independent Contrasts

When PGLS was first described by Grafen (1989), he described the method as a generalisation of Felsenstein's (1985) independent contrasts approach. At their heart, the two approaches have the same recognition of the problem of statistical non-independence of species data points as a result of shared ancestry. Independent contrasts resolves this problem by recognising that the differences ('contrasts') between closely related species or clades do provide independent data points for analyses, because they represent the outcome of independent evolutionary pathways (see Box 5.1 for details). PGLS likewise identifies from phylogeny the amount of expected correlation between species based on their shared evolutionary history, and weights for this in the generalised least squares regression calculation. Although couched in slightly different ways, ultimately, the results of PGLS, in their raw form, are the same as those derived from independent contrasts (Grafen 1989; Garland and Ives 2000; Rohlf 2001; Blomberg et al. 2012).
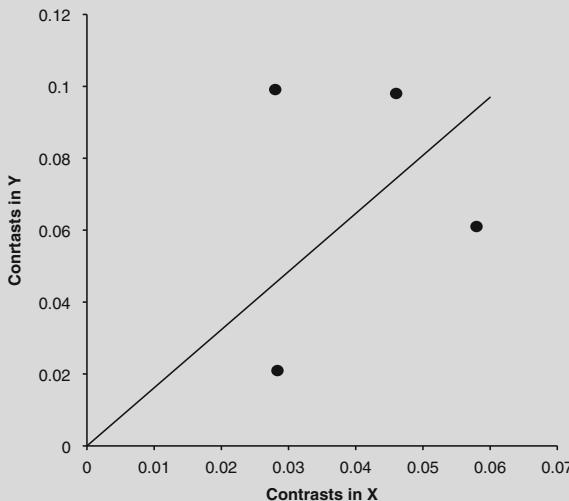

Box 5.1 Independent Contrasts

The most popular method for phylogenetic comparative analysis of continuous data has, until recently, been independent contrasts (Felsenstein 1985). The logic behind this approach is that although raw species data do not provide independent observations for analysis, differences ('contrasts') between closely related species or clades are indeed independent, because they represent the outcome of independent evolutionary pathways. By regressing the independent contrasts of one variable against the independent contrasts of another, one can estimate a regression coefficient that accounts for phylogenetic relatedness among species. Contrasts between species (or clades) are calculated downwards through the tree, with the independent variable (X) typically assigned a positive value. For the tree we discuss in this chapter (see Fig. 5.2) with 5 species, 4 independent contrasts are produced (denoted as $d1$, $d2$, $d3$, and $d4$ below).

X   1.02   1.06   0.96   0.92   0.89
Y   1.38   1.41   1.36   1.22   1.13

Trait values at nodes:
        X   Y
N1   1.04   1.395
N2   0.94   1.29
N3   0.99   1.3425

For $d1$ and $d2$, the calculation of the raw contrast values is relatively straightforward (it is just the difference between the species trait values). Calculation of the contrast values for $d3$ requires estimation of trait values for the nodes $N1$ and $N2$. These can be estimated as the means of the daughter species weighted by the daughter branch lengths to reflect amount of time over which divergence has occurred (in our example, the daughter branch lengths are the same length, so the weighted means are the same as the raw means). In order to reflect uncertainty with these estimates, the branch lengths leading to these ancestral nodes are modified by lengthening them by an amount equal to (daughter branch length $1 \times$ daughter branch length 2)/(daughter branch length $1 +$ daughter branch length 2). These modified branch lengths are shown in the brackets after the raw branch lengths on the figure. The trait values for node $N3$ can likewise be estimated as the phylogenetically weighted mean of the estimated trait values at nodes $N1$ and $N2$, and the raw contrast $d4$ subsequently calculated. As before, the branch length between the base of tree and $N3$ must be lengthened using the formula above, using the (modified) daughter branch lengths. While the raw contrast values are now statistically independent, they do not conform to another statistical requirement of having been drawn from a normal distribution with the same expected variance. Hence, they must be standardised by dividing by their standard deviation: the square root of the sum of the branch lengths leading to the two taxa in the contrast (remembering to use the modified branch lengths for internal branches in the tree). For our example, the four contrasts can now be calculated:

Contrast	Raw contrasts		Standard deviation	Standardised contrasts	
	X	Y		X	Y
d1	0.04	0.03	$\sqrt{(1+1)} = \sqrt{2}$	0.028	0.021
d2	0.04	0.14	$\sqrt{(1+1)} = \sqrt{2}$	0.028	0.099
d3	0.1	0.105	$\sqrt{(1.5+1.5)} = \sqrt{3}$	0.058	0.061
d4	0.1	0.2125	$\sqrt{(3+1.75)} = \sqrt{4.75}$	0.046	0.098

These standardised contrasts can now be plotted in a normal bivariate scatterplot.

Note that for the independent contrasts, the regression line must be forced through the origin (i.e. have a zero intercept) (Garland et al. 1992). To understand why, consider that for species A, the predicted value of Y (Y_A) is

$$Y_A = b_0 + b_1 X_A$$

where b_0 is the intercept and b_1 is the slope value. Likewise, for species B

$$Y_B = b_0 + b_1 X_B$$

For the contrast $Y_A - Y_B$, therefore,

$$Y_A - Y_B = (b_0 + b_1 X_A) - (b_0 + b_1 X_B) = b_0 + b_1 X_A - b_0 - b_1 X_B$$

Notice that the intercept b_0 terms cancel out in this equation and therefore are removed from the calculation of the regression of the contrasts:

$$Y_\mathrm{A} - Y_\mathrm{B} = b_1 X_\mathrm{A} - b_1 X_\mathrm{B} = b_1 (X_\mathrm{A} - X_\mathrm{B})$$
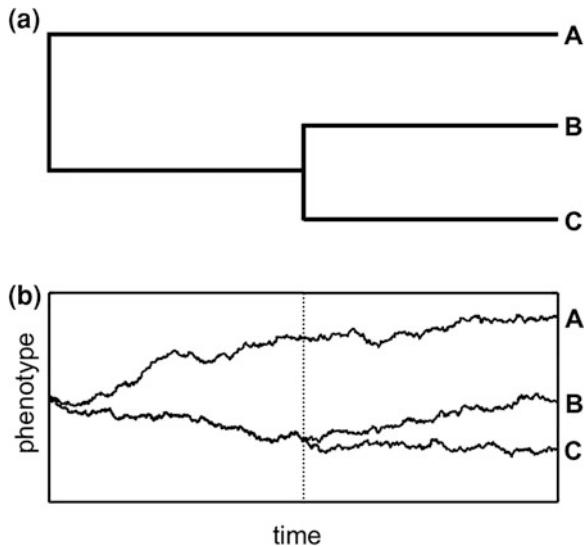
where $X_\mathrm{A} - X_\mathrm{B}$ is the contrast in X. For our example, the regression coefficient for the standardised contrasts of Y on X is 1.616.

In practice, however, most statistical packages for PGLS have an advantage over those that employ independent contrasts, because they do not automatically rely on the assumption that closely related species will necessarily be similar because of their shared phylogenetic history. In their most basic formulation, both methods assume that continuous traits evolve according to a random walk process, i.e. Brownian motion, such that the change in the value of a trait over a given period of time is given by a random number drawn from a normal distribution with a given standard deviation and mean of 0 (i.e. the value is equally likely to go up or down). Under this model, species that share a more recent common ancestor should have more similar trait values than more distantly related species because their traits have had less time to diverge (see Fig. 5.1).

However, there are many situations in which traits are evolutionarily labile, where closely related species are not necessarily more similar (Blomberg et al. 2003). Criticism that phylogenetic comparative methods might 'over-correct' for phylogeny when applied in such circumstances has been levelled for some years (e.g. Westoby et al. 1995; Björklund 1997; Rheindt et al. 2004; see also Chap. 14). In some circumstances, therefore, a traditional non-phylogenetically controlled analysis might be statistically more appropriate, not least if phylogenies are in extreme error (Abouheif 1998; Symonds 2002; Blomberg et al. 2012). Proposed solutions include presenting the results of both non-phylogenetic and phylogenetic analyses, but this does not resolve the issue of which analysis to base inference on, and it is unclear how one should proceed should the analyses produce conflicting results (see Freckleton 2009; and 'Misconceptions, problems, and pitfalls' later). Additionally, this still presents results based on two very contrasting scenarios—one which assumes no phylogenetic effect on the data and the other which assumes a strong effect. In many cases, the true effect of phylogeny is intermediate, in which case, both types of analysis would be invalid.

This problem can be overcome with PGLS, because it allows one to incorporate information on the extent of phylogenetic signal in the data (see 'Incorporating phylogenetic signal into PGLS' later). If there is no phylogenetic signal in the data, then PGLS will return estimates identical to an ordinary least squares regression analysis. If phylogenetic signal is intermediate, then PGLS can correct for phylogeny to the appropriate degree. While independent contrasts can also be adapted to deal with this issue (as in fact Felsenstein explicitly flagged in his original 1985 paper), in practice, the statistical packages which calculate independent contrasts do not automatically do so and therefore assume that the phylogeny does

Fig. 5.1 a Three-species phylogeny and **b** illustration of possible phenotypic divergence over time (i.e. evolutionary history) in those three species by a Brownian motion model of evolution. Note how the traits gradually diverge such that, typically, species *B* is most similar to *C*. Figure reproduced from Revell et al. (2008) with permission of Liam Revell and Oxford University Press

accurately describe the error structure in the data (i.e. the way species values deviate from least squares regression line—closely related species having similar errors).

PGLS and independent contrasts also present their output in slightly different ways. PGLS calculates an intercept value in the regression equation, whereas independent contrasts force the intercept through the origin (see Box 5.1 and Garland et al. 1992) and the intercept must be subsequently deduced by noting that the line goes through the phylogenetic mean (the estimated ancestral value for the response variable at the root of the phylogeny). Plots of independent contrasts also differ from plots of PGLS (which present the actual species values, rather than contrasts: see 'How to present a PGLS analysis' below). That said, contrast plots can be very informative for detecting outlier clades that are strongly influencing regression estimates.

5.2 Requirements for a PGLS Analysis

The two requirements for a PGLS analyses are a set of comparative species data and a phylogeny for those species. Chapters 2 and 3 provide greater discussion on preparing phylogenies for comparative analysis, but we provide here a quick reminder. The phylogeny may be produced de novo from phylogenetic analysis of DNA sequence data, for example. Alternatively, it may be taken from an already published source and pruned to the relevant species, or augmented as a composite

phylogeny using other sources. The phylogeny should ideally include branch lengths and be fully resolved. If not fully resolved, then some determination must be made as to whether the polytomies (when more than two species descend from a node) represent known or unknown phylogeny (i.e. the true evolutionary process—in which case, we call them 'hard' polytomies—or just uncertainty about the true pattern of relationships—'soft' polytomies). We shall discuss later (in Misconceptions, problems, and pitfalls) methods for dealing with polytomies in PGLS.

It may be that no branch length information is available for the phylogeny, in which case, one may either set all branch lengths as equal (Purvis et al. 1994), or use an algorithm such as that used by Grafen (1989) where the depth of each node in the tree is related to the number of daughter species derived from that node (see also Pagel 1992, for an alternative approach). Once compiled, the phylogeny should be formatted so that it can be read by the computer package being used for analysis. Typically, this will be a Nexus file with the stored tree presented in that file in Newick format (Maddison et al. 1997). Trees can be saved in this format by most phylogenetic analysis and tree manipulation packages.

There are several computing packages that perform PGLS: COMPARE (Martins 2004) is an online interface that will conduct PGLS and other functions including independent contrasts, but users should note that COMPARE is no longer being supported or updated. BayesTraits (Pagel and Meade 2013) implements PGLS through its package Continuous (Pagel 1997; Pagel 1999). Finally, several packages within the R statistical framework can derive PGLS estimations very quickly and efficiently, including *ape* (Paradis et al. 2004), *picante* (Kembel et al. 2010), *caper* (Orme et al. 2012), *phytools* (Revell 2012), *nlme* (Pinheiro et al. 2013), and *phyreg* (Grafen 2014).

5.3 Calculation of PGLS

5.3.1 Calculation of Parameter Estimates

The simplest way to think of PGLS is as a weighted regression. In a standard regression, each independent data point contributes equally to the estimation of the regression line. By contrast, PGLS 'downweights' points that derive from species with shared phylogenetic history. These PGLS calculations are automatically done using the appropriate statistical package (see above). Nevertheless, some knowledge of the basic approach involved in this statistical method may be informative.

In an ordinary least squares (OLS) regression model, the relationship of a response variable Y to a predictor variable X_1 can be given using the regression equation:

$$Y = b_0 + b_1 X_1 + \varepsilon \tag{5.1}$$

where b_0 is the intercept value of the regression equation, b_1 is the parameter estimate (the slope value) for the predictor, and ε is the residual error (i.e. for a given point, how far it falls off the regression line). Of course, there may also be other predictor variables in the model—X_2, X_3, etc., with associated regression slope estimates (b_2, b_3, etc.), but for simplicity, we shall focus on the simplest version of linear regression.

To illustrate our discussion, we use a simple example (Fig. 5.2). Fiddler crabs of the genus *Uca* are well known for their enlarged claws, which are used in competition between males for access to females (Crane 1975). As a sexually selected trait, we might expect these claws to show positive allometry (i.e. the parameter estimate b_1 of the regression of log(claw size) on log(body size) should be greater than 1; see Rosenberg (2002) for discussion of fiddler crab claw allometry, and Bonduriansky (2007) for explanation and analysis of the idea more generally). To test this idea, we collated data on body size (carapace breadth) and claw size (propodus length) for five species from Crane (1975). We also obtained a phylogenetic topology for the group (Rosenberg 2001).

For a simple regression with one predictor (X), the slope of the regression line $b1$ is given by

$$b_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})} \tag{5.2}$$

where n is the sample size, X_i is the ith value of X (up to the last value X_n), and $\bar{X}$ represents the mean value of X (0.97). Likewise for Y_i and $\bar{Y}$ (1.30). The intercept b_0 then simply follows:

$$b_0 = \bar{Y} - b_1 \bar{X} \tag{5.3}$$

For our fiddler crab data, the OLS estimate of the allometric equation is log(propodus length) $= -0.229 + 1.577 \times$ log(carapace breadth), with the b_1 term appearing to support the idea of positive allometry in claw length. The parameter estimates b_0, b_1, b_2, and so on (collectively denoted as the vector $\boldsymbol{\beta}$) are the values which minimise the residual variation from the least squares regression line.

For generalised least squares, we need to consider an additional element of the regression equation, in the form of the variance–covariance matrix, which represents the expected covariance structure of the residuals from the regression equation (see Appendix A for a more technical description of the mathematical formulation involved). In the case of OLS, the implicit assumption is that there is

	Log carapace breadth (X)	Log propodus length (Y)
chlorophthalmus	1.02	1.38
crassipes	1.06	1.41
inversa	0.96	1.36
sindensis	0.92	1.22
argillicola	0.89	1.13
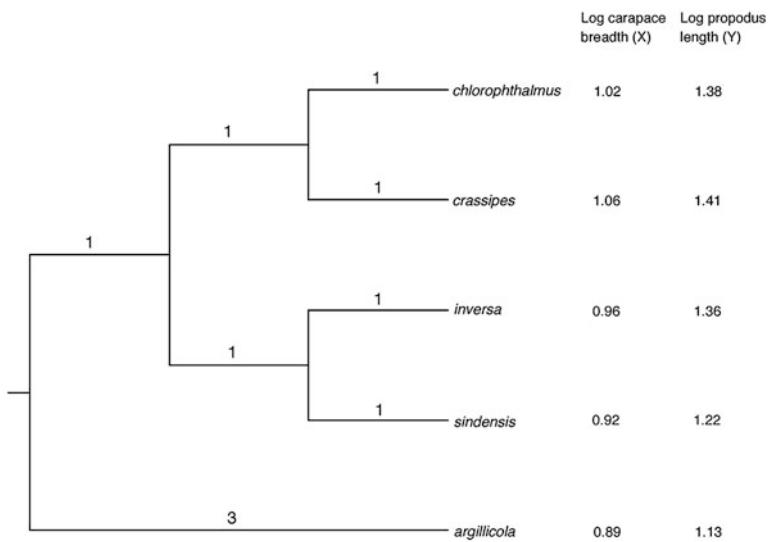
Fig. 5.2 Phylogeny of five *Uca* fiddler crab species, with morphometric data. Numbers on the phylogeny represent branch lengths

no covariance between residuals (i.e. all species are independent of each other, and residuals from closely related species are not more similar on average than residuals from distantly related species). This ($n \times n$) variance–covariance matrix is denoted as **C**, and for five species under the assumption of no phylogenetic effects on the residuals, it looks like:

$$\mathbf{C} = \begin{bmatrix} \sigma_\varepsilon^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma_\varepsilon^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_\varepsilon^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma_\varepsilon^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma_\varepsilon^2 \end{bmatrix}$$

The first row and first column represent values from comparisons with the first species (in our case *Uca chlorophthalmus*, see Fig. 5.2), the second row and column with *Uca crassipes*, and so on. Hence, the diagonal elements (the line of values from top left to bottom right) represent the variance of the residuals, while the other off-diagonal elements equal zero, meaning there is no covariation among the residuals. When this variance–covariance structure is assumed, the results of GLS are the same as those of OLS (the contribution of **C** to the regression calculation essentially drops out).

Recall that the key statistical issue with cross-species analyses is that species data points are non-independent because of their shared phylogenetic history. Consequently, the errors may also be non-independent or autocorrelated (residuals

from closely related species may be similar). Hence, there will be covariation in residuals, which we must account for in our variance–covariance matrix, **C**.

Estimation of the expected covariance structure was a key insight by Felsenstein (1973) that Grafen (1989) used in his phylogenetic regression. Like all good insights, it is elegantly simple: the expected covariance will be related to the amount of shared evolutionary history between the species. Hence, the diagonal elements (i.e. the variance elements) of the matrix are the total length of branches from the root of the tree to the tips. This will be the same for each cell if the phylogeny is ultrametric (i.e. all tips are the same distance from the root of the phylogeny), as it is in the case of our example (distance $= 3$, see Fig. 5.2). The off-diagonal covariance elements represent the total shared branch length of the evolutionary history of the two species being compared. Hence, for *U. chlorophthalmus* and *U. crassipes,* we see that each species has independent (non-shared) branch lengths of 1. Conversely, the two species share 2 branch lengths in their evolutionary history back to the root of the tree. Consequently, the value entered into column 1–row 2 (and column 2–row 1) of the matrix is 2. We can repeat this for all the other species comparisons (e.g. *U. sindensis* and *U. argillicola* do not share any evolutionary history, so their expected covariance is 0) and produce the new expected variance–covariance matrix:

$$\mathbf{C_{phyl}} = \begin{bmatrix} 3 & 2 & 1 & 1 & 0 \\ 2 & 3 & 1 & 1 & 0 \\ 1 & 1 & 3 & 2 & 0 \\ 1 & 1 & 2 & 3 & 0 \\ 0 & 0 & 0 & 0 & 3 \end{bmatrix}$$

When this new version of **C** is applied to the GLS calculation (see Appendix A), we eventually end with the PGLS solution: log(propodus length) $= -0.276 + 1.616 \times$ log(carapace breadth). Note, as we said earlier, that the regression slope coefficient is the same here as derived from independent contrasts (see Box 5.1). In this case, our final PGLS regression is not so different from the OLS regression, but there can easily be circumstances where this is not the case. We can plot and compare the two regression slopes for our data (Fig. 5.3).

5.3.2 Hypothesis Testing and Goodness of Fit

After calculating the intercept and slopes using GLS, it is common to ask questions about the magnitude of these quantities. In particular, we may be interested in whether the intercept and/or slopes are significantly different from zero. A Wald *t*-test can be conducted for each parameter in the model simply by dividing the parameter estimate by its associated standard error (i.e. the square root of the estimated variance of the parameter) and then comparing the result to a standard
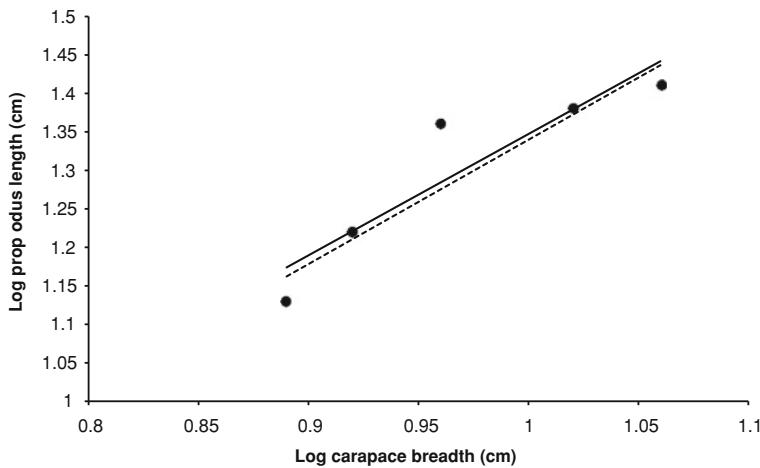
Fig. 5.3 Comparison of OLS (*solid*) and PGLS (*dashed*) regression lines for the fiddler crab claw allometry data

t distribution, using the residual degrees of freedom from the model, to calculate a P value. To test the null hypothesis that $b1 = 0$, the t statistic will therefore be

$$t = \frac{b_1}{\sqrt{\text{Var}(b_1)}} \qquad (5.4)$$

Calculation of the degrees of freedom can be non-trivial. In particular, the residual degrees of freedom may need to be reduced if there are soft polytomies in the tree (Purvis and Garland 1993; see also below). F tests for multiple variables can be similarly designed. An alternative test is the likelihood ratio chi-squared test, which has the advantage that it depends only on the likelihood of a general model (which includes the parameter) compared to a restricted model without the parameter of interest. Popular software (such as *nlme* for R) will carry out all of these procedures.

In OLS regression, it is often useful to consider how much of the total variance is explained by the model using the coefficient of variation (R^2). Unfortunately, the OLS definition of R^2 does not carry over easily into GLS. Several definitions of 'pseudo R^2' have been proposed (Menard 2000), but none of them are correct in all situations. It is therefore important to bear this issue in mind when using R^2 for PGLS regressions. Indeed, some authors prefer not to report R^2 statistics at all (e.g. Bates 2000; Lumley 2009).

A more important issue is the estimation of effect sizes and associated confidence intervals from GLS models. The parameter estimates of slopes (for continuous predictors) and the intercept and differences between means (for categorical predictors) are the most important results of the analyses. Confidence intervals for parameters can be constructed in the usual way by multiplying the

standard deviation of the parameter estimate by 1.96 to derive the 95 % confidence interval, if the sample is large (roughly >30 residual degrees of freedom), or by relating to the t distribution if the sample is smaller.

5.4 Phylogenetic Signal

5.4.1 Phylogenetic Signal and Pagel's λ

Up to now, we have assumed that the expected phylogenetic variance–covariance matrix accurately describes the error structure of the data. In other words, we assume the phylogeny is accurate (but see 'Misconceptions, problems, and pitfalls' later) and that species trait values have evolved via a Brownian motion model of gradual evolution, with the amount of evolutionary change along a branch being proportional to the branch length. However, if the phylogeny or evolutionary model is not accurate and there is in reality less or no phylogenetic covariance in the residuals (the OLS expectation), then using the phylogeny as estimated may be inappropriate. What we need is a way of determining the extent of phylogenetic autocorrelation in the data. This can be achieved by estimating phylogenetic signal.

Phylogenetic signal is the extent to which trait values are statistically related to phylogeny. In other words, phylogenetic signal indicates the extent to which closely related species tend to resemble each other (Blomberg et al. 2003). Estimation of phylogenetic signal can provide some insight into how particular traits have evolved. Thus, traits exhibiting strong phylogenetic signal (e.g. body size and morphology; Freckleton et al. 2002) have most likely evolved by gradual changes over time (e.g. a Brownian motion model of evolution). Alternatively, traits with no phylogenetic signal (e.g. many social behaviours, Blomberg et al. 2003) may either be extremely labile (they change around very much) on the time scale of phylogeny or conversely extremely stable (they do not change at all) (Revell et al. 2008).

Our interest here lies in the application of phylogenetic signal to PGLS, so we will not provide extensive discussion of the biological significance of phylogenetic signal. For interested readers, we recommend two excellent papers on the subject of phylogenetic signal (Revell et al. 2008; Kamilar and Cooper 2013).

We shall concentrate on one of the most commonly used quantitative measures of phylogenetic signal: Pagel's λ (Pagel 1997, 1999), because this measure can be directly implemented in PGLS calculations. However, there are numerous other measures of phylogenetic signal that can be employed dependent on the statistical framework and the model of evolution assumed. Each, in some way, measures the extent to which common descent of species describes the pattern of traits across species. Examples include Moran's I (Gittleman and Kot 1990), Abouheif's test for serial independence (Abouheif 1999), Grafen's ρ (Grafen 1989), the

Ornstein-Uhlenbeck model parameter α (Martins and Hansen 1997), Hansen's phylogenetic half-time (Hansen 1997), Blomberg et al.'s K (Blomberg et al. 2003), Ives and Garland's 'a' and 'd' (Chap. 9), and Fritz and Purvis's D metric (Fritz and Purvis 2010). Some of these are compatible with the PGLS framework (e.g. Grafen's ρ). For more detailed reviews, see Blomberg and Garland (2002), Münkemüller et al. (2012), and Chaps. 9, 11 and 14.

We have already introduced the expected variance–covariance matrix, $\mathbf{C_{phyl}}$, that is calculated based on the phylogenetic relationships of the species in the analysis (see above). This is the expected covariance structure, but what is the actual covariance structure? We can estimate this for a single trait or, as is the case for PGLS, the residual errors (an important distinction as we shall see later). To get one of the individual off-diagonal elements, the covariance (cov) for a pair of species (i and j) and a given trait (X) is the product of the deviation of each species from the mean of the trait:

$$cov(X_i, X_j) = (X_i - \bar{X})(X_j - \bar{X})$$

For our fiddler crab X values (log carapace breadth), the observed matrix is

$$\mathbf{C_{obs}} = \begin{bmatrix} 0.0025 & 0.0045 & -0.0005 & -0.0025 & -0.0040 \\ 0.0045 & 0.0081 & -0.0009 & -0.0045 & -0.0072 \\ -0.0005 & -0.0009 & 0.0001 & 0.0005 & 0.0008 \\ -0.0025 & -0.0045 & 0.0005 & 0.0025 & 0.0040 \\ -0.0040 & -0.0072 & 0.0008 & 0.0040 & 0.0064 \end{bmatrix}$$

We might ask which is the better 'fit' to this $\mathbf{C_{obs}}$ matrix, $\mathbf{C_{phyl}}$, or $\mathbf{C_{non\text{-}phyl}}$? It is also possible that there is intermediate phylogenetic signal in the data. Might this be a more likely scenario? We can establish this by estimating λ, which is a multiplier of the off-diagonal elements of the expected variance–covariance matrix. If λ is less than 1, this has the effect of shortening the internal branches and extending the terminal branches of the tree (see Fig. 5.4). At its extremes, $\lambda = 0$ sets the off-diagonal elements to zero producing the non-phylogenetic covariance matrix, whereas $\lambda = 1$ is identical to the expected phylogenetic covariance matrix under a Brownian motion model of evolution. Values greater than 1 are not valid because the off-diagonal values in the covariance matrix cannot exceed the diagonals in GLS (species cannot be more similar to other species than they are to themselves).

λ is not calculated through the GLS formula itself. Rather, its value is estimated through maximum likelihood estimation. A λ value of 0 is consistent with no phylogenetic signal in the trait, whereas a value of 1 is consistent with strong phylogenetic signal. Intermediate values of λ indicate intermediate phylogenetic signal. Many of the R packages cited earlier can estimate λ for individual traits. In the case of our example, the maximum likelihood value of λ for carapace breadth is 1, and for claw length, it is 0.888.
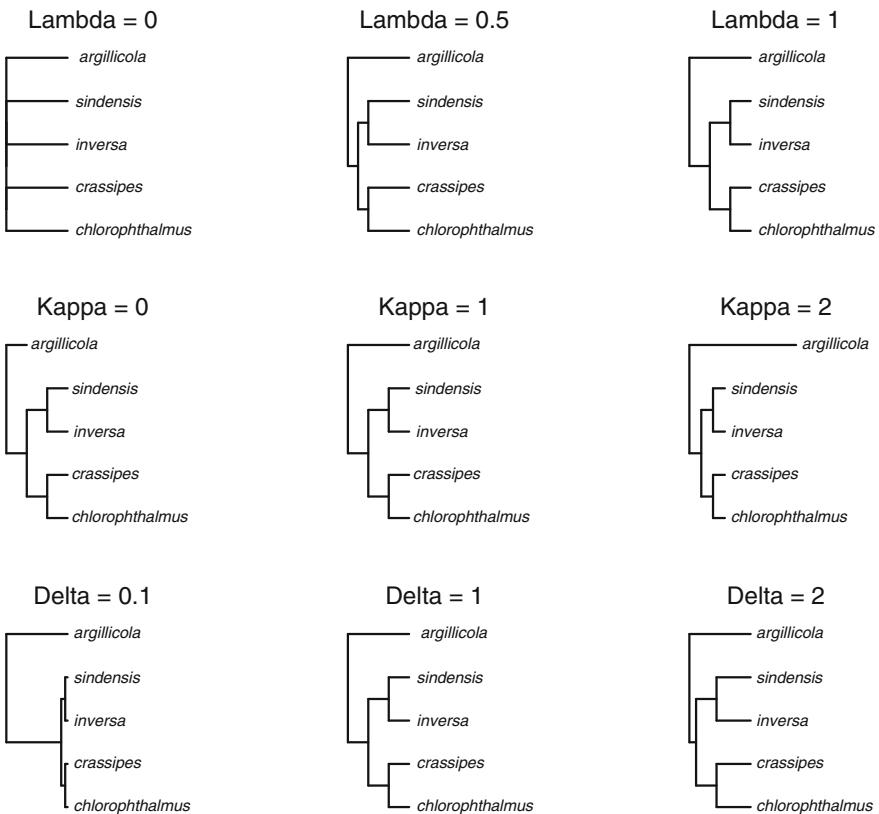
Fig. 5.4 Pagel's branch length transformations applied to the *Uca* fiddler crab phylogeny under different values of λ, δ, and κ. The $\lambda = 1$, $\delta = 1$, and $\kappa = 1$ phylogenies are identical to Fig. 5.2. Note that $\lambda = 0$ phylogeny is the same evolutionary assumption as used by traditional OLS regression (each species has independently evolved and shares no phylogenetic history)

There is no clear-cut interpretation of whether intermediate values of λ indicate 'weak' or 'strong' phylogenetic signal because it depends on the likelihood profile of λ for the specific data set (Kamilar and Cooper 2013). However, one can use likelihood ratio (LR) tests and calculate P values to assess whether the estimated maximum likelihood value of λ differs significantly from 0 or 1. As a brief aside, some authors (e.g. Pinheiro and Bates 2000) have pointed out that such likelihood ratio tests where the null value cannot exceed a certain value (such as less than 0 or more than 1) will be inherently conservative. Note that simulation studies have demonstrated that the significance of λ is also very sensitive to the number of species, and λ may perform poorly as a measure of phylogenetic signal at small sample sizes (Münkemüller et al. 2012).

It is worth pointing out that Pagel (1997, 1999) developed two other measures, related to λ, that are also branch length modifiers and are calculated through

maximum likelihood estimation. The first of these, δ, is a power transformation of the summed branch lengths from the root to the tips of the tree, and the second, κ, is a power transformation of the individual branch lengths themselves. As with λ, both can be used to infer something about the evolutionary process. δ is a measure of whether trait evolution has sped up ($\delta > 1$) or slowed down ($\delta < 1$) over evolutionary time. κ is a measure of mode of evolution, with $\kappa = 0$ depicting evolutionary change that is independent of branch length—indicating a punctuated model of evolution. Figure 5.4 illustrates the effect of different values of these parameters. As with λ, both δ and κ can also be applied to PGLS calculation (see below), although they are not as commonly utilised as λ in that context.

5.4.2 Incorporating Phylogenetic Signal into PGLS

One of the principal advantages of PGLS is that one can control for the amount of phylogenetic signal in the data by altering the properties of the variance–covariance matrix $\mathbf{C}$. In the case of independent contrasts, the usual assumption is that the phylogeny accurately describes the error structure of the data. PGLS, however, can account for intermediate levels of phylogenetic signal. With Pagel's λ, one simply multiplies the off-diagonal elements of $\mathbf{C}$ by λ and uses this new version of the matrix $\mathbf{C}_\lambda$ in the PGLS calculation. Note that the lambda multiplier can also be used to generate the modified phylogeny for use in an independent contrasts analysis, with identical results.

It is key to recognise that, in PGLS, λ applies to the residual errors from the regression model, not the strength of signal in the response variable or predictor variables. Consequently, the λ values for the PGLS regression may vary from those for the individual traits themselves (see 'Misconceptions, problems, and pitfalls' below). This is actually demonstrated by our example: the maximum likelihood estimate of λ for the regression is 0, as opposed to the individual trait values of 1 and 0.888. Thus, even though there is strong phylogenetic signal in our individual traits, there is no signal when claw length is regressed against body width, and the actual phylogenetic regression estimates will be identical to the ordinary least squares regression estimates ($b_0 = -0.229$, $b_1 = 1.577$). Note that this only applies if your phylogeny is ultrametric (all tips being the same distance from the root of the tree).

5.5 How to Present a PGLS Analysis

One advantage of PGLS is that, for the graphical presentation of relationships, one can simply plot the species data points on the relevant axes as you would do for a standard regression plot (see Fig. 5.3). The main difference is that the plot should include the PGLS regression line, rather than the standard OLS regression line.

When it comes to the presentation of the statistical analysis itself, again the presentation does not differ from what you would do with a standard regression analysis—present the PGLS estimates and standard errors, and, if appropriate, r, *t,* or F values and associated P values. The one key difference is that it is usual to present the estimate (such as λ) of the phylogenetic signal associated with the regression, along with its confidence intervals, as this provides an indication to the reader as to the extent that phylogeny is affecting the error structure of the data (remember that this is the signal associated with residual errors, not the individual variables).

Finally, because there is increasing appreciation of statistical approaches that are not based on frequentist thinking (i.e. traditional null-hypothesis significance testing with P values) (Garamszegi et al. 2009), it should be noted that PGLS is compatible with other methods of statistical inference, such as information-theoretic (e.g. using Akaike's Information Criterion) or Bayesian approaches (see Chaps. 10 and 12).

5.6 Misconceptions, Problems, and Pitfalls

As with any statistical technique, problems may arise with PGLS in practice, primarily due to violations of basic assumptions of the method. There are also several general misconceptions about phylogenetic comparative methods that apply to PGLS. For readers interested in these issues, we recommend Freckleton's (2009) review of the 'seven deadly sins of comparative analysis'. Many of these concern basic statistical assumptions, and these will be covered in the next chapter. Here, though, we address several other common practical issues.

5.6.1 Reporting Both PGLS and OLS

It is not necessary to report both PGLS and OLS (i.e. phylogenetically and non-phylogenetically controlled analyses). Unless your aim is specifically to compare the results of the two analyses (and perhaps infer the effects of phylogeny on the relationship between traits), then it is not necessary or desirable to carry out both types of analysis. The tendency to use both sets of results stemmed from concerns about the appropriateness of accounting for phylogeny in certain analyses, and perhaps a desire to 'cover one's bases' in the consequent interpretation. However, as we have seen, PGLS can explicitly take into account phylogenetic signal and hence control for it appropriately. If there is no signal in the residual structure (as we saw with our fiddler crab example), then the results of PGLS will be the same as OLS. By contrast, if there is phylogenetic signal, then PGLS will control for it, and a raw-data analysis would be statistically flawed in any event.

5.6.2 The Assumptions of the Evolutionary Model

The version of PGLS we have presented here stems from perhaps the simplest evolutionary model, the Brownian motion model (see earlier), as described by Felsenstein (1985). However, as Felsenstein (1985, p. 13) himself commented 'there are certainly many reasons for being skeptical (*sic*) of its validity'. Of course, in the absence of other knowledge, this is perhaps a reasonable starting point. However, there are other implementations of PGLS that invoke alternative evolutionary models, such as the Ornstein–Uhlenbeck model, where there is semi-random walk evolution with a tendency towards trait optima reflecting different selective regimes (see Chaps. 14 and 15; Martins and Hansen 1997; Butler and King 2004; Hansen et al. 2008). Part III of this book examines alternative evolutionary models in detail.

5.6.3 Phylogenetic Signal in the Context of PGLS

Phylogenetic signal for traits should not be used as justification for using (or not using) PGLS. As we discussed earlier, when one has a measure of phylogenetic signal for a trait, it is possible to use likelihood ratio tests to examine whether the observed value of signal differs significantly from 0 or 1. It has become quite common to argue that if one of the traits being investigated does not display any significant phylogenetic signal, then it is unnecessary to perform a phylogenetically controlled analysis (see Revell 2010 for further discussion of this issue). However, with PGLS, the assumptions regarding phylogenetic non-independence concerns the residual errors of the regression model, not the individual traits themselves. As our fiddler crab example demonstrates, it is quite possible to have strong phylogenetic signal in the traits when examined individually but not in the residual errors (and the converse is also true).

5.6.4 Dealing with Phylogenetic Inaccuracy and Uncertainty

With any phylogenetic comparative method, a fundamental assumption is that the phylogeny being used as the basis for analysis is accurate and known without error (Harvey and Pagel 1991, p. 121). Clearly, it is highly unlikely that this will be the case, and therefore, one should bear in mind that any phylogenetic comparative analysis is naturally contingent on the particular phylogeny being used. Fortunately, simulation studies have generally found that independent contrasts and PGLS are fairly robust to errors in both phylogenetic topology and branch lengths (Díaz-Uriarte and Garland 1998; Symonds 2002; Martins and Housworth 2002; Stone 2011). However, there are several points surrounding the issue of phylogenetic uncertainty that bear consideration.

First, any phylogenetic information is better than none at all (Symonds 2002). It may be that there is not a convenient single phylogeny available, in which case inference can still be based on composite trees (i.e. when phylogenetic information from several trees is fitted together), or from supertrees (Chap. 3; Bininda-Emonds 2004). Alternatively, practitioners may attempt to produce a phylogeny themselves using published DNA sequence data (e.g. from GenBank.) There are a number of phylogenetic packages available that enable use of this approach relatively quickly (e.g. phyloGenerator: Pearse and Purvis 2013). In the complete absence of any phylogenetic information or means to construct a phylogeny, the taxonomic information may suffice (indeed the original version of PGLS as described by Grafen 1989 was based around a taxonomic 'phylogeny').

Second, sometimes, there are multiple phylogenetic hypotheses for the study species, in which case the approach advocated by Harvey (1991) of conducting analyses over multiple phylogenies can be employed. For example, Symonds and Elgar (2002) demonstrated how estimation of the metabolic scaling coefficient in mammals differs depending on which of 32 phylogenies was used as the basis for analysis. Often, phylogenetic analysis itself presents hundreds of most probable trees, and it is possible to carry out PGLS using each of these phylogenetic hypotheses. De Villemereuil et al. (2012) have developed one such approach and demonstrated that by generating regression estimates across a range of candidate trees, one improves estimation of the model parameters and associated confidence intervals. Such an approach can be combined with multimodel inference (see Chap. 12). De Villemereuil et al. (2012) argue that this approach is superior to basing analysis on a single consensus tree.

Finally, one must often deal with polytomies where more than 2 branches descend from a node. These polytomies may be an actual representation of the true evolutionary branching process, or simply a lack of knowledge of that process (so-called hard and soft polytomies, respectively, Purvis and Garland 1993). Although the original formulation of PGLS (Grafen 1989) explicitly allowed for phylogenetic uncertainty in the form of polytomies, there have been ongoing issues associated with polytomies in PGLS analyses (see discussion in Rohlf 2001), including the loss of degrees of freedom in the statistical analysis. Some packages (e.g. COMPARE, Martins 2004) do not permit polytomies at all. There are three principal recommendations for dealing with polytomies in a PGLS framework. One (usually argued in the case of 'hard' polytomies) is to arbitrarily resolve the polytomies into a fully resolved bifurcating phylogeny, but to assign zero or minimal branch length (say 0.0001) to the resolved internal branches (Felsenstein 1985). The second, more appropriate for soft polytomies, is to carry out analyses on all (or at least many) possible resolutions of the phylogenetic tree in a manner analogous to the methods above for comparing across multiple phylogenies, using the Grafen (1989) algorithm to assign branch lengths (see Chap. 12). The third is simply to reduce the degrees of freedom by making them equal to 1 for soft polytomies (Purvis and Garland 1993). A final approach, based on generalised estimating equations, has also been proposed by Paradis and Claude (2002).

5.6.5 Dealing with Intraspecific Variation

In this chapter, we have considered only variation between species and therefore used species average values as our data points. Indeed, the majority of published phylogenetic comparative analyses ignore variation within species, despite its potential impact on results (see meta-analysis by Garamszegi and Møller 2010). There are methods (Chap. 7; Ives et al. 2007; Revell and Reynolds 2012) for dealing with intraspecific variation and measurement error in the PGLS framework that have been implemented in some computer packages. In short, while obtaining detailed information on intraspecific variation might not be possible for some comparative analyses, it is recommended that it be taken into account when it is possible to do so.

A.1 Appendix

A.1.1 Further Mathematical Details of the Calculation of OLS and PGLS Using Our Worked Example

An alternative way of expressing the ordinary least squares regression formula that is quicker and more effective for analysis with more than one predictor is using matrix algebra. Here, the equation to obtain regression estimates is given as

$$\boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

In this case, $\boldsymbol{\beta}$ is the vector consisting of the parameter estimates (b_0, b_1, and so on if more than one predictor variable). $\mathbf{X}$ is a matrix consisting of n rows and ($m + 1$) columns (m is the number of predictor variables), where the first column represents a constant (given the value 1 on each row), and the subsequent columns are the X values for each predictor variable. In the matrix formulation, the term $\mathbf{X}'$ denotes the 'transpose' of $\mathbf{X}$—simply put, the rows become columns, and the columns become rows.

$$\mathbf{X} = \begin{bmatrix} 1 & 1.02 \\ 1 & 1.06 \\ 1 & 0.96 \\ 1 & 0.92 \\ 1 & 0.89 \end{bmatrix} \quad \mathbf{X}' = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1.02 & 1.06 & 0.96 & 0.92 & 0.89 \end{bmatrix}$$

When multiplied together, these become $\mathbf{X'X}$, calculated as follows:

$$\mathbf{X'X} = \begin{bmatrix} 5 & 4.85 \\ 4.85 & 4.724 \end{bmatrix}$$

Here, the value in row i, column j of $\mathbf{X'X}$ equals the sum total of row i elements of $\mathbf{X'}$ multiplied by their respective column j elements of $\mathbf{X}$. So for example, row 2, column 2 of $\mathbf{X'X}$ is $(1.02 \times 1.02) + (1.06 \times 1.06) + (0.96 \times 0.96) + (0.92 \times 0.92) + (0.89 \times 0.89) = 4.724$.

Finally, the suffix $^{-1}$ applied to $\mathbf{X'X}$ indicates the 'inverse' matrix. The way the inverse matrix is calculated is somewhat complex but it is the matrix that when multiplied by it original form $(\mathbf{X'X})$ produces a matrix with 1s in the diagonal elements, and 0s in the off-diagonals (this is known as the identity matrix—see below).

$$(\mathbf{X'X})^{-1} = \begin{bmatrix} 48.21 & -49.49 \\ -49.49 & 51.02 \end{bmatrix}$$

$\mathbf{y}$ is the vector of n rows, containing the values of Y.

$$\mathbf{y} = \begin{bmatrix} 1.38 \\ 1.41 \\ 1.36 \\ 1.22 \\ 1.13 \end{bmatrix}$$

As with $\mathbf{X'X}$, for the $\mathbf{X'y}$ vector, the row i value is the overall total of each of the row i elements of $\mathbf{X'}$ multiplied by their respective counterparts in the column of $\mathbf{y}$ (i.e. row 2 = $(1.02 \times 1.38) + (1.06 \times 1.41) + (0.96 \times 1.36) + (0.92 \times 1.22) + (0.89 \times 1.13) = 6.336$.

$$\mathbf{X'y} = \begin{bmatrix} 6.5 \\ 6.336 \end{bmatrix}$$

Hence, when $(\mathbf{X'X})^{-1}$ is then multiplied by $\mathbf{X'y}$, we get the OLS solution for $\boldsymbol{\beta}$

$$\boldsymbol{\beta} = \begin{bmatrix} (48.21 \times 6.5) + (-49.49 \times 6.336) \\ (-49.49 \times 6.5) + (51.02 \times 6.336) \end{bmatrix} = \begin{bmatrix} -0.229 \\ 1.577 \end{bmatrix}$$

where the first value (-0.229) is the intercept (b_0) and the second value is the slope estimate (b_1).

For generalised least squares, an additional element is added to the regression equation, in the form of the variance–covariance matrix, which represents the expected covariance structure of the residuals from the regression equation. In the

case of OLS regression, the assumption is that there is no covariance between residuals (i.e. all species are independent of each other, and residuals from closely related species are not more similar on average than residuals from distantly related species). This ($n \times n$) variance–covariance matrix is denoted as $\mathbf{C}$, and the regression equation becomes

$$\boldsymbol{\beta} = \left(\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{C}^{-1}\mathbf{y}$$

Under the assumption that there is no covariance among the residuals and they are normally distributed, with mean = 0 and standard deviation σ_ε, then

$$\mathbf{C} = \begin{bmatrix} \sigma_\varepsilon^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma_\varepsilon^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_\varepsilon^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma_\varepsilon^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma_\varepsilon^2 \end{bmatrix}$$

The diagonal elements (the line of values from top left to bottom right) therefore represent the variance of the residuals, while the other off-diagonal elements = 0, meaning there is no covariation among the residuals. The inverse of this matrix, $\mathbf{C}^{-1}$, has essentially the same properties (all the off-diagonal elements remain as 0) except the diagonal elements now equal $1/\sigma_\varepsilon^2$. When this variance–covariance structure is assumed, the results of GLS are the same as those of OLS (the $\mathbf{C}$ part of the regression equation essentially drops out). On the other hand, if the variances are not equal, then you have a standard weighted least squares regression.

For phylogenetic generalised least squares, our expected variance–covariance matrix is $\mathbf{C_{phyl}}$ (see main text), and its inverse

$$\mathbf{C_{phyl}} = \begin{bmatrix} 3 & 2 & 1 & 1 & 0 \\ 2 & 3 & 1 & 1 & 0 \\ 1 & 1 & 3 & 2 & 0 \\ 1 & 1 & 2 & 3 & 0 \\ 0 & 0 & 0 & 0 & 3 \end{bmatrix}$$

$$\mathbf{C_{phyl}^{-1}} = \begin{bmatrix} 0.619 & -0.381 & -0.048 & -0.048 & 0 \\ -0.381 & 0.619 & -0.048 & -0.048 & 0 \\ -0.048 & -0.048 & 0.619 & -0.381 & 0 \\ -0.048 & -0.048 & -0.381 & 0.619 & 0 \\ 0 & 0 & 0 & 0 & 0.333 \end{bmatrix}$$

Taking apart the components of the GLS regression equation, we first calculate the product $\mathbf{X}'\mathbf{C}^{-1}$ whose row i and column j values are the total of the ith row of

$\mathbf{X}'$ multiplied by the jth column of $\mathbf{C}^{-1}$. So, for example, row 2, column 3 of $\mathbf{X}'\mathbf{C}^{-1}$ is $(1.02 \times -0.048) + (1.06 \times -0.048) + (0.96 \times 0.619) + (0.92 \times -0.381) + (0.89 \times 0) = 0.144$

$$\mathbf{X}'\mathbf{C}^{-1} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1.02 & 1.06 & 0.96 & 0.92 & 0.89 \end{bmatrix} \times \begin{bmatrix} 0.619 & -0.381 & -0.048 & -0.048 & 0 \\ -0.381 & 0.619 & -0.048 & -0.048 & 0 \\ -0.048 & -0.048 & 0.619 & -0.381 & 0 \\ -0.048 & -0.048 & -0.381 & 0.619 & 0 \\ 0 & 0 & 0 & 0 & 0.333 \end{bmatrix}$$

$$= \begin{bmatrix} 0.142 & 0.142 & 0.142 & 0.142 & 0.333 \\ 0.137 & 0.177 & 0.144 & 0.104 & 0.296 \end{bmatrix}$$

In similar fashion $\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}$ is therefore

$$\mathbf{X}'\mathbf{C}^{-1}\mathbf{X} = \begin{bmatrix} 0.142 & 0.142 & 0.142 & 0.142 & 0.333 \\ 0.137 & 0.177 & 0.144 & 0.104 & 0.296 \end{bmatrix} \times \begin{bmatrix} 1 & 1.02 \\ 1 & 1.06 \\ 1 & 0.96 \\ 1 & 0.92 \\ 1 & 0.89 \end{bmatrix}$$

$$= \begin{bmatrix} 0.901 & 0.859 \\ 0.859 & 0.825 \end{bmatrix}$$

The inverse of which is

$$(\mathbf{X}'\mathbf{C}^{-1}\mathbf{X})^{-1} = \begin{bmatrix} 130.141 & -135.389 \\ -135.389 & 142.060 \end{bmatrix}$$

The second component of the GLS regression equation $\mathbf{X}'\mathbf{C}^{-1}\mathbf{y}$ follows likewise as

$$\mathbf{X}'\mathbf{C}^{-1}\mathbf{y} = \begin{bmatrix} 0.142 & 0.142 & 0.142 & 0.142 & 0.142 \\ 0.137 & 0.177 & 0.144 & 0.104 & 0.296 \end{bmatrix} \times \begin{bmatrix} 1.38 \\ 1.41 \\ 1.36 \\ 1.22 \\ 1.13 \end{bmatrix} = \begin{bmatrix} 1.139 \\ 1.097 \end{bmatrix}$$

where, for example, the first row value $(1.139) = (0.142 \times 1.38) + (0.142 \times 1.41) + (0.142 \times 1.36) + (0.142 \times 1.22) + (0.142 \times 1.13)$.

Finally, we can combine our two products to obtain the PGLS solution for $\boldsymbol{\beta}$.

$$\boldsymbol{\beta}_{\mathbf{PGLS}} = (\mathbf{X}'\mathbf{C}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}^{-1}\mathbf{y} = \begin{bmatrix} 130.141 & -135.389 \\ -135.389 & 142.060 \end{bmatrix} \times \begin{bmatrix} 1.139 \\ 1.097 \end{bmatrix}$$

$$= \begin{bmatrix} -0.276 \\ 1.616 \end{bmatrix}$$

where $b_0 = -0.276$ and $b_1 = 1.616$.

References

Abouheif E (1998) Random trees and the comparative method: a cautionary tale. Evolution 52:1197–1204

Abouheif E (1999) A method for testing the assumption of phylogenetic independence in comparative data. Evol Ecol Res 1:895–909

Bates D (2000) fortunes: R fortunes. R package version 1.5-0, http://CRAN.R-project.org/package=fortunes

Bininda-Emonds ORP (ed) (2004) Phylogenetic supertrees: combining information to reveal the tree of life. Kluwer Academic Publishers, Dordrecht

Björklund M (1997) Are 'comparative methods' always necessary? Oikos 80:607–612

Blomberg SP, Garland T Jr (2002) Tempo and mode in evolution: phylogenetic inertia, adaptation and comparative methods. J Evol Biol 15:899–910

Blomberg SP, Garland T Jr, Ives AR (2003) Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution 57:717–745

Blomberg SP, Lefevre JG, Wells JA, Waterhouse M (2012) Independent contrasts and PGLS regression estimators are equivalent. Syst Biol 61:382–391

Bonduriansky R (2007) Sexual selection and allometry: a critical reappraisal of the evidence and ideas. Evolution 61:838–849

Butler MA, King AA (2004) Phylogenetic comparative analysis: a modelling approach for adaptive evolution. Am Nat 164:683–695

Crane J (1975) Fiddler crabs of the Wworld: ocypodidae: genus Uca. Princeton University Press, Princeton

De Villemereuil P, Wells JA, Edwards RD, Blomberg SP (2012) Bayesian models for comparative analysis integrating phylogenetic uncertainty. BMC Evol Biol 12:102

Díaz-Uriarte R, Garland T Jr (1998) Effects of branch lengths errors on the performance of phylogenetically independent contrasts. Syst Biol 47:654–672

Felsenstein J (1973) Maximum-likelihood estimation of evolutionary trees from continuous characters. Am J Human Genet 25:471–492

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

Freckleton RP (2009) The seven deadly sins of comparative analysis. J Evol Biol 22:1367–1375

Freckleton RP, Harvey PH, Pagel M (2002) Phylogenetic analysis and comparative data: a test and review of evidence. Am Nat 160:712–726

Fritz SA, Purvis A (2010) Selectivity in mammalian extinction risk and threat types: a new measure of phylogenetic signal strength in binary traits. Conserv Biol 24:1042–1051

Garamszegi LZ, Møller AP (2010) Effects of sample size and intraspecific variation in phylogenetic comparative studies: a meta-analytic review. Biol Rev 85:797–805

Garamszegi LZ, Calhim S, Dochtermann N, Hegyi G, Hurd PL, Jørgensen C, Kutsukake N, Lajeunesse MJ, Pollard KA, Schielzeth H, Symonds MRE, Nakagawa S (2009) Changing philosophies and tools for statistical inferences in behavioral ecology. Behav Ecol 20:1363–1375

Garland T Jr, Ives AR (2000) Using the past to predict the present: confidence intervals for regression equations in phylogenetic comparative methods. Am Nat 155:346–364

Garland T Jr, Harvey PH, Ives AR (1992) Procedures for the analysis of comparative data using phylogenetically independent contrasts. Syst Biol 41:18–32

Gittleman JL, Kot M (1990) Adaptation: statistics and a null model for estimating phylogenetic effects. Syst Zool 39:227–241

Grafen A (1989) The phylogenetic regression. Phil Trans R Soc B 326:119–157

Grafen A (2014) phyreg: Implements the phylogenetic regression of Grafen (1989). http://cran.r-project.org/web/packages/phyreg/index.html

Hansen TF (1997) Stabilizing selection and the comparative analysis of adaptation. Evolution 51:1341–1351

Hansen TF, Pienaar J, Orzack SH (2008) A comparative method for studying adaptation to a randomly evolving environment. Evolution 62:1965–1977

Harvey PH (1991) Comparing uncertain relationships: the Swedes in revolt. Trends Ecol Evol 6:38–39

Harvey PH, Pagel MD (1991) The comparative method in evolutionary biology. Oxford University Press, Oxford

Huey RB (1987) Phylogeny, history and the comparative method. In: Feder ME, Bennett AF, Burggren WW, Huey RB (eds) New directions in ecological physiology. Cambridge University Press, Cambridge, pp 76–101

Ives AR, Garland T Jr (2010) Phylogenetic logistic regression for binary dependent variables. Syst Biol 59:9–26

Ives AR, Midford PE, Garland T Jr (2007) Within-species variation and measurement error in phylogenetic comparative methods. Syst Biol 56:252–270

Kamilar JM, Cooper N (2013) Phylogenetic signal in primate behaviour, ecology and life history. Phil Trans R Soc B 368:20120341

Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP, Webb CO (2010) Picante: R tools for integrating phylogenies and ecology. Bioinformatics 26:1463–1464

Lumley T (2009) fortunes: R fortunes. R package version 1.5-0, http://CRAN.R-project.org/package=fortunes

Losos JB (2011) Seeing the forest for the trees: the limitations of phylogenies in comparative biology. Am Nat 177:709–727

Maddison DR, Swofford DL, Maddison WP (1997) Nexus: an extensible file format for systematic information. Syst Biol 46:590–621

Maddison WP (1990) A method for testing the correlated evolution of two binary characters: are gains or losses concentrated on certain branches of a phylogenetic trees? Evolution 44:539–557

Maddison WP (2000) Testing character correlation using pairwise comparisons on a phylogeny. J Theor Biol 202:195–204

Martins EP (2004) COMPARE. Version 4.6b. Computer programs for the statistical analysis of comparative data. Department of Biology, Indiana University, Bloomington. http://compare.bio.indiana.edu/

Martins EP, Hansen TF (1997) Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. Am Nat 149:646–667

Martins EP, Housworth EA (2002) Phylogeny shape and the phylogenetic comparative method. Syst Biol 51:873–880

Menard S (2000) Coefficients of determination for multiple logistic regression analysis. Am Stat 54(1):17–24

Münkemüller T, Lavergne S, Bzeznik B, Dray S, Jombart T, Schiffers K, Thuiller W (2012) How to measure and test phylogenetic signal. Methods Ecol Evol 3:743–756

Orme D, Freckleton R, Thomas G, Petzoldt T, Fritz S, Isaac N, Pearse W (2012) caper: comparative analysis of phylogenetics and evolution in R. http://CRAN.R-project.org/package=caper

Pagel MD (1992) A method for the analysis of comparative data. J Theor Biol 156:431–442

Pagel M (1994) Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proc R Soc B 255:37–45

Pagel M (1997) Inferring evolutionary processes from phylogenies. Zool Scripta 26:331–348

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

Pagel M, Meade A (2013) BayesTraits version 2.0 (Beta). University of Reading. http://www.evolution.rdg.ac.uk/BayesTraits.html

Paradis E, Claude J (2002) Analysis of comparative data using generalized estimating equations. J Theor Biol 218:175–185

Paradis E, Claude J, Strimmer K (2004) APE: analysis of phylogenetics and evolution in R language. Bioinformatics 20:289–290

Pearse WD, Purvis A (2013) phyloGenerator: an automated phylogeny generation tool for ecologists. Methods Ecol Evol 4:692–698

Pinheiro JC, Bates DM (2000) Mixes-effects models in S and S-PLUS. Springer, Berlin

Pinheiro J, Bates D, DebRoy S, Sarker D, R Development Core Team (2013) nlme: linear and nonlinear mixed effects models. R package version 3.1-111. http://cran.r-project.org/web/packages/nlme/index.html

Promislow DEL, Harvey PH (1990) Living fast and dying young: a comparative analysis of life-history variation among mammals. J Zool 220:417–437

Purvis A, Garland T Jr (1993) Polytomies in comparative analysis of continuous characters. Syst Biol 42:569–575

Purvis A, Gittleman JL, Luh H-K (1994) Truth or consequences: effects of phylogenetic accuracy on two comparative methods. J Theor Biol 167:293–300

Revell LJ (2010) Phylogenetic signal and linear regression on species data. Methods Ecol Evol 1:319–329

Revell LJ (2012) phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol Evol 3:217–223

Revell LJ, Reynolds RG (2012) A new Bayesian method for fitting evolutionary models to comparative data with intraspecific variation. Evolution 66:2697–2707

Revell LJ, Harmon LJ, Collar DC (2008) Phylogenetic signal, evolutionary process and rate. Syst Biol 57:591–601

Rheindt FE, Grafe TU, Abouheif E (2004) Rapidly evolving traits and the comparative method: how important is testing for phylogenetic signal? Evol Ecol Res 6:377–396

Ridley M (1983) The explanation of organic diversity. Oxford University Press, Oxford

Rohlf FJ (2001) Comparative methods for the analysis of continuous variables: geometric interpretations. Evolution 55:2143–2160

Rosenberg MS (2001) The systematics and taxonomy of fiddler crabs: a phylogeny of the genus Uca. J Crust Biol 21:839–869

Rosenberg MS (2002) Fiddler crab claw shape variation: a geometric morphometric analysis across the genus Uca (Crustacea: Brachyura: Ocypodidae). Biol J Linn Soc 75:147–162

Stone EA (2011) Why the phylogenetic regression appears robust to tree misspecification. Syst Biol 60:245–260

Symonds MRE (2002) The effects of topological inaccuracy in evolutionary trees on the phylogenetic comparative method of independent contrasts. Syst Biol 51:541–553

Symonds MRE, Elgar MA (2002) Phylogeny affects estimation of metabolic scaling in mammals. Evolution 56:2330–2333

Westoby M, Leishman MR, Lord JM (1995) On misinterpreting the 'phylogenetic correction'. J Ecol 83:531–534

Chapter 6
Statistical Issues and Assumptions of Phylogenetic Generalized Least Squares

Roger Mundry

Abstract Using phylogenetic generalized least squares (PGLS) means to fit a linear regression aiming to investigate the impact of one or several predictor variables on a single response variable while controlling for potential phylogenetic signal in the response (and, hence, non-independence of the residuals). The key difference between PGLS and standard (multiple) regression is that PGLS allows us to control for residuals being potentially non-independent due to the phylogenetic history of the taxa investigated. While the assumptions of PGLS regarding the underlying processes of evolution and the correlation of the predictor and response variables with the phylogeny have received considerable attention, much less focus has been put on the checks of model reliability and stability commonly used in case of standard general linear models. However, several of these checks could be similarly applied in the context of PGLS as well. Here, I describe how such checks of model stability and reliability could be applied in the context of a PGLS and what could be done in case they reveal potential problems. Besides treating general questions regarding the conceptual and technical validity of the model, I consider issues regarding the sample size, collinearity among the predictors, the distribution of the predictors and the residuals, model stability, and drawing inference based on P-values. Finally, I emphasize the need for reporting checks of assumptions (and their results) in publications.

6.1 Introduction

The method of phylogenetic generalized least squares (PGLS) is an extension of the general linear model. The general linear model, in turn, is a unified framework allowing us to analyze the impact of one or several predictor variables on a single

R. Mundry (✉)
Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
e-mail: roger_mundry@eva.mpg.de

quantitative and continuous response (e.g., Quinn and Keough 2002). In fact, it is nothing else than the classical multiple regression. Categorical predictors ('factors') can be easily included into multiple regression using (usually dummy) coding, interactions can be modeled by including products of predictors into the model, and nonlinear effects are usually included by including transformed predictors (in addition to or instead of the untransformed ones) into the model (for more details about how the effects of factors, interactions, and nonlinear effects can be modeled, see below and, e.g., Cohen and Cohen 1983; Aiken and West 1991). Hence, the general linear model encompasses multiple regression, ANOVA, ANCOVA, and the t-tests.

A crucial assumption of the general linear model is independence of the residuals. This assumption is likely to be violated when the cases in the data set (see Glossary for a definition of some of the terms used here) represent different taxa (e.g., species) which share larger or smaller fractions of their evolutionary history (Felsenstein 1985). Obviously, taxa sharing a larger fraction of their evolutionary history (i.e., having a more recent common ancestor) are likely to be more similar to one another (even after considering the effects of various potential predictors) leading to non-independent residuals.

The method of PGLS (Grafen 1989) has been developed to cope with such phylogenetically driven non-independent residuals. PGLS is an extension of the general linear model, allowing us to account for the phylogenetic history and, by this means, controlling for potential non-independence of the data and leading to independent residuals. The properties and assumptions of PGLS with regard to the assumed evolutionary process and its consequences for the character states of the taxa investigated have received quite some attention. For instance, different models of character evolution (e.g., Brownian motion, Ornstein–Uhlenbeck; Felsenstein 1985, 1988; Hansen 1997; Pagel 1999; Chaps. 5 and 15) can be assumed, or different branch-length scaling parameters (e.g., lambda, kappa, delta; Pagel 1999; Chap. 5) can be chosen, and the particular choices obviously can have crucial implications for the results of the analysis (e.g., Díaz-Uriarte and Garland 1996, 1998; Martins et al. 2002). Similarly, the details of the estimation process can have clear impacts on the reliability of the results (e.g., Revell 2010), as could heterogeneities of the underlying process across the clade investigated (Garland and Ives 2000), heterogeneous sampling (Freckleton and Jetz 2009; Cooper et al. 2010), and errors in the phylogenetic tree (e.g., Díaz-Uriarte and Garland 1998), and these issues (to mention just a few) need to be carefully addressed when using PGLS. Here, I focus on issues and assumptions of PGLS that arise from its similarity with multiple regression. First of all, a PGLS makes assumptions about the distribution of the residuals that are largely identical to those of multiple regression. Furthermore, a model's reliability depends on the distribution(s) of the predictor(s), their number in relation to the sample size, as well as absence of strong collinearity and influential cases. A violation of the assumptions about the residuals or model instability can severely affect the conclusions drawn, and hence, it is of crucial importance that these are thoroughly checked and an assessment is made about how much the models can be trusted.

In the following, I shall treat the statistical assumptions of PGLS and also questions regarding model validity, stability, and reliability. I shall begin with questions regarding the conceptual and technical validity of the model and subsequently consider issues related to the number and distribution of the predictors and interrelations among them (i.e., issues that could be dealt with before the model is fitted). Following that, I shall consider assumptions about the residuals and questions related to model stability (i.e., issues that can be dealt with only after the model was fitted). Finally, I shall briefly touch on questions related to drawing inference based on significance testing and also give recommendations regarding the reporting of the analysis. Most sections have complementary parts in the Online Practical Material (http://www.mpcm-evolution.com) where I show how the respective checks can be conducted in R (Core Team 2013).

It must be stated here that the majority of the issues I consider are not specifically linked to a particular statistical approach (i.e., whether inference is drawn based on information theory, null-hypothesis significance testing or in a Bayesian framework; the exception is the section about drawing inference using null-hypothesis significance testing). Instead, they are generic in the sense that regardless of which particular statistical philosophy one follows, one should consider them. In this context, it might be worth noting that the issues I consider are also not specific to phylogenetic analyses but generic to general linear models. In fact, what I present here are the issues I regularly consider when fitting linear models (such as standard multiple regressions, generalized linear models, or linear mixed models). I also want to emphasize that most of the issues I consider here are not really 'assumptions' of PGLS (or linear models in general) in the sense that a model requires them to be fulfilled (as is the case with assumptions about the residuals). In fact, linear models do not rely on assumptions such as absence of influential cases or collinearity or a certain sample size in relation to the number of cases. However, the confidence one can have in the conclusions drawn from a model might crucially depend on what investigations about, for instance, model stability reveal.

A crucial assumption of all statistical analyses is that the data are correct and complete and for predictors that they are measured without error. A special issue in the context of phylogenetic data is that data availability could vary systematically with species traits (see Garamszegi and Møller 2012). I take it for granted here that the data are correct and complete and that missing data occur at random. Furthermore, I want to emphasize that I am solely focusing on the *statistical* issues related to the use of PGLS and not assumptions regarding the validity of the entire approach, for instance, assumptions about the particular model of evolution (e.g., Brownian motion or Ornstein–Uhlenbeck; Chaps. 5 and 15), the parameter used to model phylogenetic signal in the residuals (e.g., lambda or kappa; see Revell 2010; Chap. 5), and the correctness of the phylogeny used. These 'phylogenetic' assumptions have been treated in quite some detail elsewhere (see above and also Harvey and Pagel 1991 or Nunn 2011, and for possible approaches to deal with uncertainty in the phylogeny or the model of evolution, see, e.g., Chap. 12). However, to my knowledge, with the exception of the distribution of the residuals

(Freckleton 2009), statistical issues like model stability which have relevance for a PGLS as in any other general linear model have received less attention so far.

6.2 Model Development

This section deals with the general outline of the model. In fact, before any model can be fitted, it is required to think about which particular terms should be included and how these should enter the model. In fact, fitting a model crucially requires considering whether the model makes any sense at all or, in other words, whether it is appropriate for the question considered. This is what this section is about. Specifically, it briefly touches the question which predictors at all and also whether interactions and/or nonlinear terms (or interactions involving nonlinear terms) should be considered. I also briefly consider some technical questions regarding model validity that particularly come into play when interactions and/or nonlinear terms are included.

6.2.1 Conceptual Validity of the Model

The first issue to be considered is which predictors should be included in the model. I sometimes have the impression that many researchers believe that this question is not really an issue anymore since 'model selection' provides a simple and automated approach telling which predictors are important or not. However, it must be clearly stated that model selection and significance testing are two approaches to statistical inference that are not conformable (e.g., Burnham and Anderson 2002; Mundry 2011), and once model selection has been applied, significance tests are meaningless. Hence, whenever inference should be based on null-hypothesis significance testing (i.e., P-values), decisions about which predictors are to be included in the model have to be based on scientific reasoning and cannot be substituted by an automated model selection approach. Moreover, also the proponents of model selection clearly emphasize the need of a careful development of the models to be fitted (e.g., Burnham and Anderson 2002). Of course, decisions about which particular predictors to be considered or controlled for are highly specific to each individual investigation. Nevertheless, I am convinced that many considerations are straightforward in this context. For instance, it seems obvious that any investigation of the impact of whichever predictors on longevity or brain size *must* control for body size. I am convinced that every researcher can easily come up with other such examples in her or his own research area.

Pretty much the same applies to decisions about which interactions to be considered in a model.[1] Again, such decisions *must* be made based on reasoning and cannot be solved using technical solutions such as 'exploratory data analysis' at least when inference should be based on *P*-values. The reason for this is the exact same as for individual predictors. Hence, decisions about which interactions are to be included into a model must be made prior to any analysis and based on reasoning, and they should completely disregard the actual data at hand. As for individual predictors, such decisions are highly specific to the individual study, but again, common sense seems to potentially provide a lot in this context, too. For instance, when one wants to investigate the impact of environmental complexity (e.g., biodiversity) on (relative) brain size, one might need to consider the interaction between environmental complexity and diet as a predictor, simply because folivorous species (that potentially could eat pretty much everything that is 'green') might face much less difficulties in finding food than frugivorous or carnivorous species (which presumably are more specialized in their dietary needs), and this difference might be particularly pronounced in complex environments with high species richness.

The same line of reasoning applies to nonlinear terms. Most commonly nonlinear terms are included as squared terms allowing for an optimum regarding the impact of a covariate on the response (in the sense that one allows for the response to show particularly large (or small) values at intermediate values of the covariate). The details of reasoning about which covariates should be considered to potentially have nonlinear effects are, of course, again very specific to the particular investigation. However, as before, common sense and reasoning can presumably reveal clear hints about which covariates should be considered to be included as nonlinear terms. For instance, in a study of the impact of group size (the predictor) on brain size (the response) within a certain clade (e.g., genus or family), one might hypothesize that intermediate group sizes lead to particularly large relative brain sizes, because small groups are socially not very complex by definition and large groups could be socially not very complex because of being anonymous.

A final issue to be considered here is that adding terms to the model potentially conflicts with the size of the data set which might impose limitations on model complexity (see below). However, my personal preference when being in such a conflict is to give priority to the 'right' model, which means to potentially have more terms than desired in the model. The simple reason is that from a model which is known to be wrong, for instance, because a potentially important confounder or a likely interaction is neglected, potentially not much can be learned (actually, such a model would violate the assumption of independent residuals; see below). I also quite frequently had the impression that a neglected confounder, interaction, etc., can lead to an inflated error variance making tests actually

---

[1] Having an interaction between two predictors in a model means to allow for a situation where the impact of one of the two on the response is dependent on the value or state of the other and vice versa. Interactions can involve two or more covariates, two or more factors, and any mixture of covariates and factors.

conservative, and this effect can be stronger than the reduction in power and model stability coming along with including an additional term to the model.

6.2.2 Technical Validity of the Model

Whenever an interaction and/or nonlinear (e.g., squared) term is in a model, this necessitates the terms encompassed by them to be in the model as well. Practically, this means that when a two-way interaction is in a model, then the two terms interacting (the 'main effects') *must* be in the model as well (e.g., Aiken and West 1991). Correspondingly, when a model comprises a three-way interaction, the model must also comprise all three two-way interactions encompassed by the three-way interaction and also the respective three main effects. Similarly, when a squared covariate is in a model, this *must* comprise also the respective unsquared covariate. If these requirements are not fulfilled, the model is actually meaningless, and the results revealed for the respective interaction or squared term have no interpretation. It might seem trivial to state this, but among published papers, one finds a surprisingly large proportion being unclear about this point.

6.2.3 Scaling and Centering of the Predictors

A question frequently arising is whether covariates should be z-transformed to a mean of zero and a standard deviation of one (note that a z-transformation is always done *after* a potential other transformation of a covariate; see below). Strictly spoken, a z-transformation is never really required, but it might make interpretation easier quite frequently. First of all, the coefficients obtained for covariates being z-transformed represent the average change in the response per standard deviation of the covariate, and hence, they are directly comparable between all covariates, regardless of what they present and on which scale they were measured (Aiken and West 1991). Hence, getting comparable coefficients is one of the main reasons for z-transforming covariates. The other main reason for z-transforming covariates is to enhance model interpretability in case of models including interactions and/or nonlinear (usually squared) terms (Schielzeth 2010). The reason is that when an interaction between two covariates is in a model, then the coefficients derived for the respective main effects indicate their effect at the respective *other* covariate having a value of zero. Many covariates, however, never have a value of zero in nature (e.g., brain size, body size, life span). As an example, if an interaction between the effects of brain size and life span on a response is in a model, the estimated coefficients would be response $= c_0 + c_1 \times$ brain size $+ c_2 \times$ life span $+ c_3 \times$ brain size $\times$ life span. The coefficient c_1 then represents the effect of brain size on the response for life span being zero—not a very meaningful quantity.

If brain size and life span are z-transformed (meaning that their average is zero), the coefficients c_1 and c_2 have a much more reasonable interpretation; that is, they indicate the effect of the two covariates at the *average* of the respective other covariate. Pretty much the same logic applies whenever nonlinear terms are in a model. For the same reasons, one could also consider centering manually dummy coded factors to a mean of zero (see Schielzeth 2010 for a more in-depth account on these considerations). A final reason to z-transform covariates and scale dummy coded factors is to easier create plots of the modeled effects of individual predictors on the response (because ignoring all other terms in the model when plotting the particular effect implies assuming them to be at their average).

However, besides these many advantages of z-transforming all covariates (and potentially also to center all dummy coded factors), it has the disadvantage that coefficients reported for different data sets are not comparable anymore since the standard deviation of any particular covariate will vary between data sets and studies (and a coefficient obtained for a z-transformed covariate indicates the change of the response per *standard deviation* of the covariate). As a consequence, one should routinely report the original standard deviations of the covariates being z-transformed (and also their means; Schielzeth 2010).

6.3 Statistical Reliability of the Model

This section deals with preparatory steps potentially taken to avoid certain problems as well as assumptions about the residuals and questions regarding model stability. If these reveal problems, the validity of the conclusions might be questionable for solely statistical reasons.

6.3.1 Things to be Checked Before the Model Is Fitted

A couple of issues can (and should) be dealt with prior to fitting any model. These refer to the number of predictors in relation to the number of cases, the distribution of the predictors, and absence of strong collinearity.

6.3.1.1 Number of Predictors and Sample Size

Maybe the first and simplest to check is the number of cases (i.e., sample size, N) in relation to the number of predictors (k).[2] In fact, for standard linear models other

---

[2] Note that 'number of predictors' should actually be labeled 'number of estimated terms' (meaning that a factor would be counted as the number of its levels minus 1, interactions and

than PGLS, it is well established that the sample size should considerably exceed the number of predictors. If this is not the case, the power of the analysis decreases (potentially considerably), and the results are likely to suffer from instability (i.e., slight changes in the data may lead to drastic changes in the results). However, no simple universally accepted and applicable rule for what would be an appropriate ratio of sample size to the number of predictors does exist (see Field (2005) for recommendations regarding multiple regression). This makes sense, though, since any consideration of sample size needs to take into account expected (or minimum detectable) effect sizes and the power desired (e.g., Cohen 1988; Gelman and Hill 2007). Since expected and minimum effect sizes to be detected are rarely available (if at all in phylogenetic and many other analyses), I tend to use a very simple rule which is that the number of cases should be roughly and at least 10 times the number of estimated terms (including the intercept and, e.g., lambda in case of a PGLS). Surely, this rule is extremely crude and overly simplistic, and whenever possible, one should replace it by something more appropriate (ideally a power analysis based on simulations of the expected or minimum effect size to be detected, conducted using phylogenetic data as close as possible to the ones eventually to be analyzed). At the same time, though, the ten-cases-per-estimated-term rule is simple, allows for a rapid exclusion of model–data combinations that do not make much sense at all (e.g., 5 predictors and 10 cases), and is better than nothing.

If the number of predictors is too large (identified by whichever rule), one needs to reduce them. At least three options do exist in such a case: (1) exclude predictors based on reasoning about which are the least likely to be of relevance for the response under question; (2) conduct a principal component or factor analysis and use the derived (principal component or factor) scores rather than the original covariates (for more about principal component and factor analysis, see, e.g., Budaev 2010, and for phylogenetic principal components analysis see Revell 2009 and Polly et al. 2013); and (3) exclude predictors based on collinearity (i.e., the variance inflation factors revealed for them; see section about collinearity). In the context of phylogenetic analyses, where the number of available taxa might be limited, one will at occasions be confronted with a situation where the model seems to be too complex for the size of the available data set. It is hard to give general recommendations about what can be done in such a situation. However, as stated above, I tend to give priority to the 'right' model (with regard to the predictors included) over one that is oversimplified only to meet an assumed limit of model complexity. After all, the model used needs to be appropriate with regard to the hypotheses to be addressed and the variables to be controlled for. However, a model being too complex might appear unstable (see below). On the other hand, though, a model with two or three predictors might still reveal reasonable results for surprisingly small data sets (with, e.g., just some 15 cases).

(Footnote 2 continued)

squared terms need to be considered, and in the context of a PGLS a parameter like lambda needs to be counted as well).

6.3.1.2 Distribution of Quantitative Predictors (Covariates)

Presumably, the far majority of phylogenetic data sets comprise at least one 'covariate,' i.e., a quantitative predictor such as average group size, body size, brain size, or longevity. Perhaps surprisingly, a PGLS (like a general linear model) does not make any explicit or direct assumptions about the distribution of such covariates (Quinn and Keough 2002; Field 2005). Nevertheless, it is good practice to generally inspect the distribution of each covariate (using, e.g., a histogram or qq-plot; see below) before fitting the model. Besides the fact that such an inspection can reveal typos, it can give hints to potentially 'problematic' (i.e., influential) cases (see below). Such problematic cases are more likely to arise with skewed covariates (Quinn and Keough 2002; Field 2005). For instance, when a covariate is right-skewed (see Glossary), then the values at the upper (right) end of its distribution are likely to have more influence on the model than those in the middle or at the lower end (simply because there are fewer large than small values; Quinn and Keough 2002). As a consequence, one routinely should check the distributions of covariates and try to transform those that are skewed, trying to achieve a roughly symmetrical distribution (e.g., roughly normal or uniform). Most commonly, a log- or square root transformation can be used for this purpose. The log-transformation is 'stronger' and requires all values to be positive, and the square root transformation requires all values to be nonnegative (see also Fig. 6.1 and Box 6.1). One needs to keep in mind, though, that a PGLS ultimately does not make any particular assumption about the distribution of a covariate, and hence, even if a covariate is quite skewed, it might be that a model with the original, untransformed, covariate is more appropriate. However, most usually, it is a good idea to transform skewed covariates right away.

Box 6.1 log-transform or other model?

I quite frequently encountered the belief that rather than transforming a predictor (and/or a response), one should run a modified (i.e., 'nonlinear') model. While this is frequently required when the assumed function linking the response with the predictors is more complex (e.g., sigmoidal), there are many situations where a simple transformation of the predictor (e.g., the log or square root) seems to be the by far easiest and also a fully appropriate solution (Quinn and Keough 2002). This is frequently the case because of the 'logarithmicity' of life. What I mean by this is that the relevance and per-ception of variables in biological systems are frequently (perhaps usually) logarithmic. For instance, the exact same evolutionary change in body mass of, say, 10 kg would be probably considered 'huge' in case of a mouse or bird, 'minor' in case of an antelope, and 'negligible' in case of an elephant. Correspondingly, one would expect traits that covary with body mass to be affected by the relative and not the absolute change in body mass. Such an expectation can be easily accounted for by taking the logarithm and not the

absolute value of body mass as a predictor into the model. Practically, after log-transforming a variable, two species that differ by the same *relative* value differ by the same quantity in the transformed variable. For instance, the difference in log-transformed weights being 100 and 10 would be the same as that between weights being 1,000 and 100. By the way, the base of the logarithm (e.g., 2, 10 or e) does not matter, but probably the natural logarithm (base e) is most commonly used.

6.3.1.3 Categorical Predictors (Factors)

Categorical predictors (i.e., predictors representing a quality, such as herbivorous, frugivorous, or carnivorous) are usually (and by default) entered into PGLS models by dummy coding them, and after this, factors are modeled as any 'normal' (i.e., quantitative) predictor (Cohen and Cohen 1983; Aiken and West 1991). As is obvious, dummy coded categorical predictors do not have the property of having a 'distribution' of a particular shape (e.g., normal of uniform). Nevertheless, it is important to inspect the frequency 'distribution' of a factor with particular focus on the frequencies of the rarer levels. Specifically, none of the levels should be too rare. From a more statistical perspective, cases of rare levels are likely to be unduly influential as compared to cases of more common levels. But also common sense tells that rare levels are unlikely to reveal much reasonable information. Assume, for instance, a study of male song complexity as a function of male investment in the brood in songbirds. Such a study might need to control for factors like whether the species is a cooperative breeder or whether females and males engage in duets (both partners regularly singing a structured song pattern simultaneously). However, both cooperative breeding and duetting species might be very rare in the data set (say, each is represented by less than five species). As a consequence, it might be a better idea to drop those species from the data rather than including two additional factors into the model. The argument for doing so would be that from such small numbers of species, not much can be learned about the respective factors anyways.

When several factors are relevant for the model, it can also be important to check how many times the combinations of their levels do occur in the data set (particularly when their interactions should be included). Here, pretty much the same logic as for the frequency distribution of the levels of a single factor applies: If a combination of levels of two or more factors is very rare (using the above example: if there were, e.g., only two species being cooperative breeders *and* duetting), then such cases are particularly likely to be relatively influential. This becomes even more of an issue when the interaction between the two factors should be included (and one should keep in mind that including the interaction between two factors is only possible and makes sense when all combinations of all
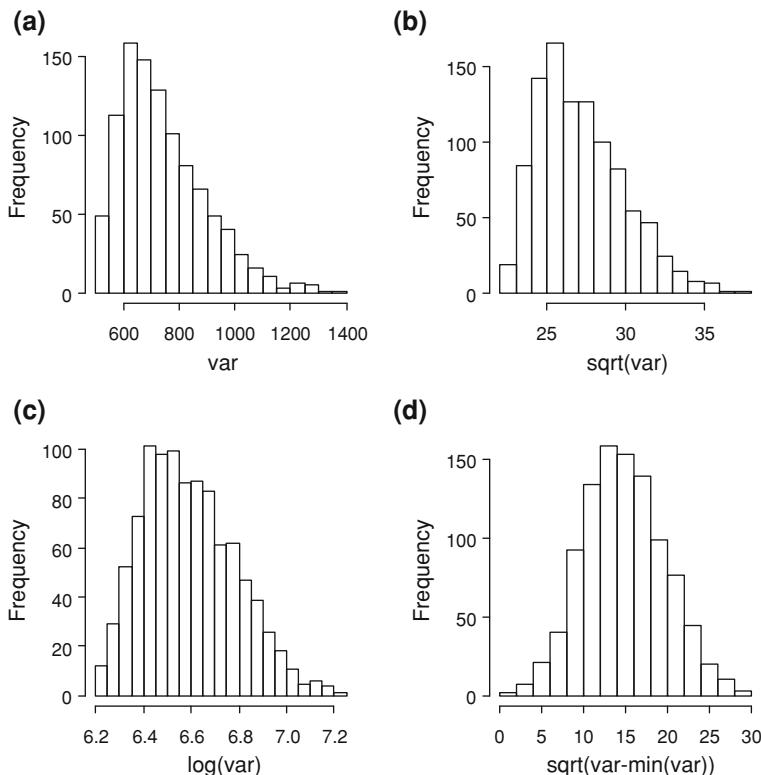
Fig. 6.1 Illustration of the effects of transformations. The original variable **a** is moderately right-skewed with a minimum considerably larger than zero. Neither a square root **b** nor a log-transformation **c** is very effective in removing the skew. However, a standardization of the variable to a minimum of zero and a subsequent square root transformation remove the skew very effectively **d**

their levels occur at least twice in the data[3]). However, such rare levels or combinations of levels of factors might mainly compromise model stability with regard to the factors themselves but not much for the effects of other predictors. An investigation of model stability (see below) will reveal whether this is the case.

6.3.1.4 Collinearity

Absence of strong collinearity (a.k.a., 'multicollinearity') among the predictors is an important requirement for the validity of the results of linear models (Quinn and

---

[3] For estimating the effect of an interaction reasonably well, more cases per combination of the levels of the factors would be needed.

Keough 2002; Field 2005). 'Absence of collinearity among predictors' basically means that they are not redundant or, in other words, that each of them provides information which is not given by the others (or combinations of them). The simplest case of collinearity is just two predictors being pretty correlated. However, absence of any large (absolute) correlations between the predictors does not rule out collinearity. For instance, one might think of a model which includes the average number of males and females per social group (of the taxa considered) *and* sex ratio or total group size. The latter two predictors are derived from the former two and, hence, do not provide any additional information. As a consequence, any set of predictors comprising three of the variables mentioned would be highly collinear. Finally, collinearity can also arise from the number of predictors being too large.

The consequences of collinearity are simple. First, conclusions about the impact of individual predictors being collinear with others get (potentially very) unreliable and uncertain (Quinn and Keough 2002; Field 2005; Zuur et al. 2010). This manifests in increased standard errors of parameter estimates (and consequently larger confidence intervals), and particularly, non-significance (i.e., $P > 0.05$) can be largely a consequence of collinearity rather than being indicative of absence of an effect. Second, a model suffering from collinearity is likely to be (potentially very) unstable, meaning that small changes in the data can lead to (potentially drastic) changes in the parameter estimates obtained for the collinear predictors (see Freckleton (2011) for an in-depth treatment of the effects of collinearity).

A simple mean to detect collinearity is inspection of so-called variance inflation factors (VIF; Quinn and Keough 2002; Field 2005). These are based on the following principle: For each predictor, one model is fitted, taking the particular predictor as the response and the others as the predictors (note that this means that one gets one VIF value for each of the predictors; note also that a response variable is not needed in this context). Then, the R^2 of the respective model is calculated and this, in turn, is used to derive the VIF, as $VIF = 1/(1-R^2)$ (see Fox and Monette 1992 for how factors with more than two levels are treated in this context). As is obvious, when the R^2 gets large, the denominator of the equation approaches zero, and hence, the VIF gets large, too. An 'ideal' VIF, i.e., one indicating no collinearity whatsoever at all, has a value of one, and the larger the VIF the worse. For the question of what is a too large VIF, no simple answer does exist. Occasionally, one sees the rule that a value of ten or larger is clearly indicative of a problem (Quinn and Keough 2002; Field 2005) but also much smaller thresholds of three or two have been suggested as being indicative of potential collinearity issues (Zuur et al. 2010).

Unfortunately, collinearity is likely to arise in phylogenetic analyses since many of the commonly used predictors are most likely to scale allometrically (i.e., correlate with body size), leading to potentially conflicting needs of controlling for body size and avoiding collinearity. In case clear collinearity is detected, one could consider the simple omission of one or several of the predictors associated with a large VIF (based on the reasoning that they anyway do not provide much information in addition to that provided by the other predictors, but see Freckleton
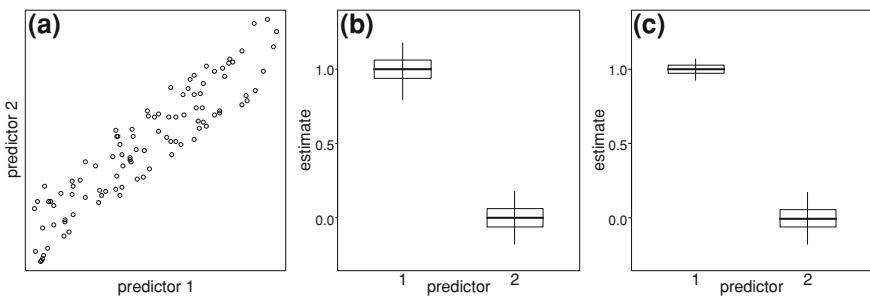
Fig. 6.2 Two collinear predictors that each show reasonable variation at a given value of the respective other predictor (**a**). When sampling predictor 1 from a uniform distribution with a minimum of 0 and a maximum of 10, and predictor 2 as the value of predictor 1 plus a value sampled from a uniform distribution with a minimum of 0 and a maximum of 4 (with a total $N = 100$), the average variance inflation factor for both of them (across 1,000 simulation) was 7.40. When simulating a response by adding a value from a standard normal distribution (mean = 0, sd = 1) to predictor 1, it appeared that the effects of the two predictors could still be assessed quite reliably ((**b**); median, quartiles, and percentiles 2.5 and 97.5 % of 1,000 simulated data sets). However, when predictor 2 was simulated independently of predictor 1, the average variance inflation factor dropped to 1.01 and the variance in the estimate for predictor 1 decreased (**c**). The simulation was based on non-phylogenetic data and a standard linear model

2011). An alternative is to combine predictors (Freckleton 2011) using, for instance, a principal component or factor analysis and use the derived (principal component or factor) scores rather than the original covariates (e.g., Quinn and Keough 2002; for more about principal component or factor analysis, see Budaev 2010). Another option is to try to include more and/or other taxa in the data set, selecting them such that collinearity among the predictors in the phylogeny is minimized, an approach basically being a modification of the method of 'phylogenetic targeting' (Arnold and Nunn 2010) applying another criterion.

However, some level of collinearity will frequently be unavoidable in phylogenetic data sets and the question arises of what to do when this is the case. First of all, it is worth noting that collinearity affects estimation of the effects of predictors being collinear with one another but not that of others. Hence, if predictors merely in the model to control for their effects are collinear to one another but the key 'test predictors' (see below) are not collinear with others, collinearity might be less of a reason to worry. Secondly, even a larger VIF associated with one of the test predictors is not necessarily 'deadly.' This is particularly the case when there is reasonable variation in the predictor at given values of the other(s) it is collinear with, meaning that their independent effects can still be assessed with some certainty. However, estimation would still be more precise when there was no collinearity (Fig. 6.2; see also Freckleton 2011). If one is uncertain about whether collinearity affects conclusions (having VIF values between, say, two and ten), one should assess model stability. This could be done by comparing results from a model with all the predictors included with those obtained from additional models excluding one or several of the collinear predictors and checking whether this has

larger consequences for the conclusions drawn. If this is the case, it basically reflects insufficient knowledge and an inability to tease apart the effects of the collinear predictors.

As mentioned earlier, one can use VIF values to exclude predictors in case their number is too large in relation to the number of cases (see above). The procedure is simple: One fits models one after the other and iteratively excludes the predictor with the largest VIF.

6.3.2 Things to be Checked After the Model was Fitted

After the model structure being clarified and having completed the initial checks of the data, one finally can fit the model. Once this is done, one needs to worry about two further issues, namely the distribution of the residuals and absence of influential cases.

6.3.2.1 Distribution of the Residuals (And the Response)

Since the PGLS is, in essence, a general linear model (i.e., a multiple regression) accounting for non-independence of the residuals arising from a phylogenetic history, it has the same assumptions about the distribution of the residuals as a standard general linear model (Freckleton 2009). In particular, these are normality and homogeneity of the residuals.

Normality of the Residuals

Normality of the residuals[4] is a somewhat pretty critical assumption, although, to my knowledge, the consequences of its violation have rarely been systematically investigated in the framework of a PGLS. Taking the general linear model as a reference, it appears that the consequences of violations of this assumption depend on the particular pattern of the violation and also the sample size. For instance, when the sample size is larger and the residuals are somewhat symmetrically distributed, type I and type II error rates are not that strongly affected (see, e.g., Zuur et al. 2010 and references therein), and for PGLS, Grafen and Ridley (1996) showed that it performed reasonably well even when the response was actually binary (however, one should better consider approaches specifically developed for binary responses; see Ives and Garland 2010; Chap. 9). However, skewed

---

[4] Note that residuals of a PGLS are actually multivariate normal (Freckleton et al. 2011), which has implications for practical checks of their distribution; see the Online Practical Material (http://www.mpcm-evolution.com) for more.

distributions of the residuals and particularly residual distributions comprising one or a few outliers are a reason to worry (and these usually lead to a reduction of power). Hence, one should routinely inspect the distribution of the residuals. Most usually, one will employ visual checks of the assumption, namely a qq-plot (and potentially also a simple histogram). If these reveal residuals to be skewed, one should consider transformations of the predictor(s) and/or the response (see also next section). In case the residuals appear to comprise outliers, one should first check whether these result from coding errors, and if this is not given, one should ask oneself whether there are any predictors missing (e.g., data usually arising from both sexes but occasionally from only females or males). When there appears a missing predictor, one should consider including it into the model or dropping rare levels in case it being a factor (see above). Note that an outlier in the residuals might also give hint to an evolutionary singularity (Nunn 2011; Chap. 21).

A frequently asked question is whether one should use formal checks of the distribution of residuals (e.g., test for normality). However, a *P*-value is always a function of (at least) two properties of the data: the effect size (here, the deviation from normality) and the sample size. Practically, this means that the *P*-value alone does not provide a simple criterion for rejection (or not) of the assumption that the residuals are normally distributed but can only be interpreted in conjunction with the sample size. Hence, eyeballing a qq-plot (or a histogram) is usually considered the most reliable and valuable tool for assessing whether the residuals are roughly normally distributed (e.g., Quinn and Keough; Zuur et al. 2010).

Homogeneity of the Residuals

The other, and presumably more critical, assumption about the residuals is that they are homogeneous (a.k.a. 'homoskedastic'). This means that the variation in the residuals should be the same, regardless of the particular constellation of values of the predictors. Not much is known about consequences of violations of this assumption (i.e., 'heteroskedasticity') in the framework of a PGLS. However, the consequences of heterogeneous residuals are probably not such that they simply lead to increased type I or type II error rates. For instance, for the independent-samples t-test (which is a special case of the general linear model), heteroskedasticity can lead to clearly increased type I *and* type II error rates, depending on whether residual variance correlates positively or negatively with sample size (e.g., Ramsey 1980). As a consequence, one should be quite vary regarding violations of this assumption.

A check of this assumption is pretty straightforward, and again, one usually employs a visual check (Quinn and Keough 2002; see also above). What one usually does is plotting the residuals of the model (on the y-axis) against its fitted values (on the x-axis), and what one wants to see here is nothing (i.e., no pattern in the cloud of points). More specifically, what should be discernible from this plot is simply no pattern whatsoever at all; that is, the residuals should show the same pattern of scatter around zero over the entire range of the fitted values. Figure 6.3 shows an example of homogeneous residuals (a) and also two common patterns of
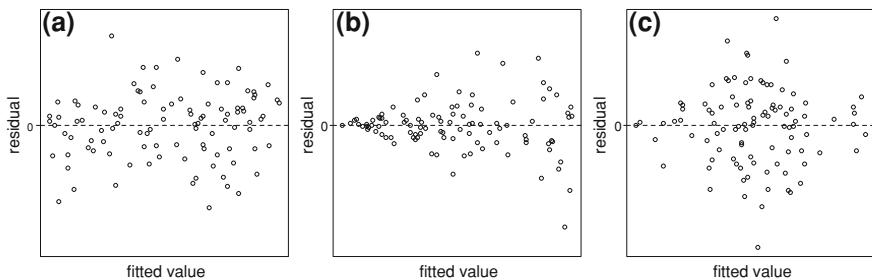
Fig. 6.3 Illustration of the assumption of homogeneous residuals and deviations from it. When the residuals are homogeneous, no relation between residual variance and the fitted value is discernible (**a**). The probably most common violation of the assumption of homoskedasticity reveals a residual variance being positively correlated with the fitted value (**b**). In such a case, a log- or square root transformation of the response might alleviate the issue. Quite frequently, one also sees a pattern where the residual variance is large at intermediate fitted values and small at small and large fitted values (**c**). Such a pattern can occur as a consequence of bottom and ceiling effects in the response and/or when one or several predictors show many intermediate and few small and large values. When the response is bounded between zero and one, the arcsine of the square root-transformed response may (or may not) alleviate the issue; otherwise, a careful selection of the taxa included in the model may help. In case of such a pattern, one should also be particularly worried about influential cases

the assumption being violated (b, c). In certain cases of heteroskedasticity (i.e., when the residual variance is positively or negatively correlated with the fitted value), a transformation of the response might alleviate the issue (Fig. 6.3). The question might arise as to how such plots usually look like when the assumption of homogeneous residuals is not violated. In the Online Practical Material (http://www.mpcm-evolution.com) I show how experience regarding the issue can be rapidly gained.

A potential cause of heterogeneous residuals is a misspecified model. For instance, when a covariate has a nonlinear effect on the response which is not accounted for by the model, then this might become obvious from a plot of residuals against fitted values (Fig. 6.4). Heterogeneous residuals could also arise when the pattern of impact of the predictor on the response is not homogeneous over the entire phylogeny (Rohlf 2006) or when an important main effect or interaction is missing. Obviously, the results of a model with such unmodelled structure in the residuals can be pretty misleading. Hence, one should try to include the missing terms (or revise the taxa investigated in case the effects of a predictor being heterogeneous across the phylogeny). However, such an a posteriori change in the model structure might need to be accounted for when it comes to inference (since a hypothesis being generated based on inspection of some data and then tested using the same data will lead to biased standard errors, confidence intervals, and P-values; Chatfield 1995).
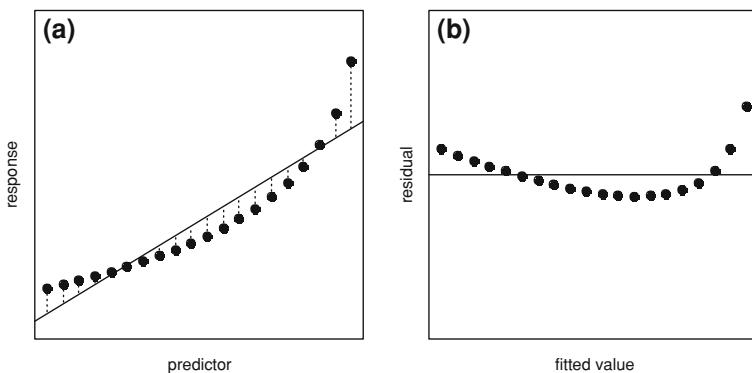
Fig. 6.4 Example of a misspecified model leading to a clear pattern in the residuals. Here, the predictor has a nonlinear effect on the response not being accounted for by the model (*straight line* in **a**). As a consequence, the residuals are large for small and large fitted values and small for intermediate fitted values (**b**)

6.3.2.2 Absence of Influential Cases

Another issue that can only be evaluated once the model was fitted is absence of influential cases. Absence of influential cases means that there are no data points that are particularly influential for the results of the model, meaning that the results do not largely differ depending on whether or not the particular cases are or are not in the data. The influence of a particular data point on the results can be investigated from various perspectives, and which of them is relevant depends on the focus of the analysis. In case of a PGLS where inference about particular effects is usually crucial for the results, the estimates of the effects (and perhaps their standard errors) will usually be in the focus of considerations of model stability.

In the framework of a standard general linear model or generalized linear model, model stability is usually investigated by simply removing cases one at a time and checking how much the model changes, whereby 'model change' is usually investigated in terms of estimated coefficients and fitted values ('dfbetas' and 'dffits,' respectively; Field 2005). For the assessment of whether any particular case is considered too influential (in the sense of questioning the results of the model), no simple cutoffs do exist, and even if it were possible to unambiguously identify influential cases, the question of what to do with such piece of information is not easy to be answered. Assume, for instance, a situation where an analysis of a data set with 100 taxa revealed an estimate for the effect of a certain covariate being 1. Dropping taxa one by one from the data set revealed that for 99 of the taxa removed, the estimate under consideration had values between 0.9 and 1.1, but when removing the remainder taxon, the estimate obtained was -0.2. What would one conclude from such a result? It is tempting to conclude that 99 out of 100 data sets revealed an estimate within a small range (0.9–1.1) indicating a stable result. However, it is more the opposite which is true here. In fact, this particular outcome also implies that the actual estimate revealed crucially depends on a single taxon

being in the data set or not (in the sense that if it is not included, the results are totally different). But what would then consider to be 'truth'? Obviously, there is no simple answer to these questions and the probably most sensitive one can do is to ask oneself why the particular taxon being found to be so influential is so different from the others? A simple (and not the most implausible explanation) is the existence of an error (of coding, a typo, etc.), but potentially, it is also due to a variable missing in the model or an evolutionary singularity (Nunn 2011; Chap. 21). In fact, the taxon might crucially differ from all others in the data, for instance, by being the only cooperative breeder among the species (to use the example from above). In such a case, one would probably conclude that the story might be different for cooperative breeders and exclude them from the analysis because of poor coverage of this factor by the data at hand (but note that such operations should not be required when the appropriate model is properly specified in advance and when factors and covariates are properly investigated in advance; see above).

6.4 Drawing Conclusions

The final step after the model was developed and fitted and decided to be reliable and stable is to draw conclusions from the results. This penultimate section deals with some issues coming along with this step.

6.4.1 Full-Null Model Comparison

A very frequently neglected issue in the context of the interpretation of more complex models (i.e., models with more than a single predictor) is multiple testing. In fact, as shown by Forstmeier and Schielzeth (2011), each individual term in a model has a five percent chance of revealing significance in the absence of any effects. As a consequence, models with more than a single predictor have an increased chance of at least one of them revealing erroneous significance. A simple means to protect from this increased probability of false positives is to conduct a *full-null model comparison*. The rationale behind this is as follows: The full model is simply the model fitted, and the null model is a model comprising only the intercept (but see below). Comparing the two gives an overall *P*-value for the impact of the predictors as a whole (which accounts for the number of predictors), and if this reveals significance, one can conclude that among the set of predictors being in the model, there is at least one having a significant impact on the response.[5] The *P*-values obtained for the individual predictors are then considered

---

[5] Note that this requires the model to be fitted using maximum likelihood; see the Online Practical Material (http://www.mpcm-evolution.com) for more details.

in a fashion similar to post hoc comparisons; that is, they are considered significant only if the full-null model comparison revealed significance.

In this context, it might be helpful to distinguish between *test predictors* and *control predictors*. This distinction can be helpful when not all predictors being in the full model are of interest in the light of the hypotheses to be investigated, but some are included only to control for their potential effects. Such predictors could and should be kept in the null model to target the full-null model comparison at the test predictors. For instance, assume an investigation of the impact of social and environmental complexity on brain size. In such a case, body size will be only in the model to control for the obviously trivial effect of body size on brain size. Hence, one could and should keep body size in the null model since its effect is trivial and taken for granted and not of any particular relevance for the hypotheses to be investigated (except for that it needs to be controlled for). I want to emphasize, though, that the entire exercise of the full-null model comparison is only of relevance when one takes a frequentist's perspective on inference, that is, when one draws conclusions based on *P*-values.

6.4.2 Inference About Individual Terms

There are two final issues with regard to drawing inference based on *P*-values that deserve attention here. The first concerns inference about factors with more than two levels being represented by more than a single term in the model. As explained above, factors are usually dummy coded, revealing one dummy variable for each level of the factor except the reference level. Hence, a factor having, for instance, three levels will appear in the model output with two terms. As a consequence, one does not get an overall *P*-value for the effect of the factor, but two *P*-values, each testing the difference between the respective dummy coded level and the reference level. However, quite frequently, an overall *P*-value of the effect of the factor as a whole will be desired (particularly since the actual *P*-values shown are pretty much a random selection out of the possible ones since the reference category is frequently just the one which is the alphanumerically first). Such a test can be obtained using the same logic as for the full-null model comparison: A reduced model lacking the factor but comprising all other terms present in the full model is fitted and then compared with the full model using an F-test. If this test reveals significance, the factor under consideration has a significant impact on the response (see, e.g., Cohen and Cohen 1983 for details).

The other concerns inference of terms involved in interactions, for which interpretation must be made in light of the interaction they are involved in. Assume, for instance, a main effect involved in a two-way interaction: Its estimates (and the *P*-value associated with it) are conditional on the particular value of the other main effect it is interacting with. This can be seen when one considers the model equation with regard to the effects of the two predictors ('A' and 'B')

interacting with one another, namely response $= c_0 + c_1 \times A + c_2 \times B + c_3 \times A \times B$. Since the effect of A on the response is represented by two terms in the model, one of which (c_3) modeling how its effect depends on the value of the other, c_1 models the effect of predictor A at predictor B having a value of zero (which would be its average if B were a z-transformed covariate or the reference category if B were a dummy coded factor). The same logic applies when it comes to the interpretation of two-way interactions involved in a three-way interaction and so on and also when it comes to the interpretation of linear effects involved in nonlinear effects like squared terms. This limited (precisely conditional) interpretation of terms involved in an interaction (or a nonlinear term) needs to be kept in mind when interpreting the results (see Schielzeth 2010 for a detailed account on the issue).

6.5 Transparency of the Analysis

A proper analysis requires a number of decisions. These include (but are not limited to) decisions about which main effects, interactions, and nonlinear terms are to be included in the model, potential transformations of predictors and/or the response, and subsequent z-transformations of the covariates. Furthermore, conducting a proper analysis means to conduct several checks of model validity and stability which might reveal good or bad results or something in between. All these decisions as well as the results of checks of model stability and validity are an integral and important part of the analysis and should be reported in the respective paper. Finally, also, the software used for the analysis should be mentioned (and in case of R being used also the key functions and the packages (including the version number) that provided them). Otherwise, the reader will be unable to judge the reliability of the analysis and cannot know to what extend the results can be trusted (see also Freckleton 2009). From my understanding, this means to (1) thoroughly outline the reasoning that took place when the model was formulated; (2) clearly formulate the model analyzed; (3) clearly describe the preparatory steps taken (e.g., variable transformations); and (4) clearly describe the steps taken to evaluate the model's assumptions and stability and what they revealed. Nowadays, there is usually the option to put extensive information into supplementary materials made available online, and we should use this option! Only by providing fully transparent analyses, we make our projects repeatable, and repeatability is at the core of science.

6.6 Concluding Remarks

What I have presented here are some of the steps that should be taken to verify the statistical validity and reliability of a PGLS model. I focused on issues and assumptions frequently checked for standard general linear models which begin

with the formulation of a scientifically and logically adequate model and continue with its technically valid implementation. At the core of the establishment of the statistical validity and reliability of a model are issues regarding the number of estimated terms in relation to the number of cases, the distribution(s) of the predictor(s), absence of collinearity and influential cases, and assumptions about the residuals. Finally, drawing inference requires special care, and all the steps conducted to verify the model need a proper documentation. What I presented here are largely the steps taken to validate the reliability of a general linear model that I more or less simply transferred to PGLS. However, besides the steps I considered here, the validity of a PGLS model also most crucially depends on a variety of issues specific to PGLS (see introduction) and these must not be neglected neither.

With this chapter, I hope to create more attention for evaluations of whether the assumptions of a given model are fulfilled and to what extend a given model is stable and reliable. Only, once the assumptions of a model and its stability have been checked carefully, one can know how confident one can be about its results.

Glossary

Case	Set of entries in the data referring to the same taxon; represented by one row in the data set and corresponds to one tip in the phylogeny.
Covariate	Quantitative predictor variable.
Dummy coding	Way of representing a factor in a linear model, by turning it into a set of 'quantitative' variables. One level of the factor is defined the 'reference' level (or reference category), and for each of the other levels a variable is created which is one if the respective case in the data set is of that level and zero otherwise. The estimate derived for a dummy coded variable reveals the degree by which the response in the coded level differs from that of the reference level.

Factor	Qualitative (or categorical) predictor variable.
General linear model	Unified approach to test the effect(s) of one or several quantitative or categorical predictors on a single quantitative response; makes the assumptions of normally and homogeneously distributed residuals; multiple regression, ANOVA, ANCOVA, and the t-tests are all just special cases of the general linear model.
Level	Particular value of a factor (for instance, the factor 'sex' has the levels 'female' and 'male').
Predictor (variable)	Variable for which its influence on the response variable should be investigated or controlled for; can be a factor or a covariate.
Response (variable)	Variable being in the focus of the study and for which it should be investigated how one or several predictors influence it.
Right (left) skewed distribution	Distribution with many small and few large values (a left skewed distribution shows the opposite pattern).

References

Aiken LS, West SG (1991) Multiple regression: testing and interpreting interactions. Sage, Newbury Park

Arnold C, Nunn CL (2010) Phylogenetic targeting of research effort in evolutionary biology. Am Nat 176:601–612

Budaev SV (2010) Using principal components and factor analysis in animal behaviour research: caveats and guidelines. Ethology 116:472–480

Burnham KP, Anderson DR (2002) Model selection and multimodel inference, 2nd edn. Springer, Berlin

Chatfield C (1995) Model uncertainty, data mining and statistical inference. J Roy Stat Soc A 158:419–466

Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd edn. Lawrence Erlbaum Associates, New York

Cohen J, Cohen P (1983) Applied multiple regression/correlation analysis for the behavioral sciences, 2nd edn. Lawrence Erlbaum Associates Inc., New Jersey

Cooper N, Jetz W, Freckleton RP (2010) Phylogenetic comparative approaches for studying niche conservatism. J Evol Biol 23:2529–2539

Díaz-Uriarte R, Garland T Jr (1996) Testing hypotheses of correlated evolution using phylogenetically independent contrasts: sensitivity to deviations from Brownian motion. Syst Biol 45:27–47

Díaz-Uriarte R, Garland T Jr (1998) Effects of branch length errors on the performance of phylogenetically independent contrasts. Syst Biol 47:654–672

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

Felsenstein J (1988) Phylogenies and quantitative characters. Ann Rev Ecol Syst 19:445–471

Field A (2005) Discovering statistics using SPSS. Sage Publications, London

Forstmeier W, Schielzeth H (2011) Cryptic multiple hypotheses testing in linear models: overestimated effect sizes and the winner's curse. Behav Ecol Sociobiol 65:47–55

Fox J, Monette G (1992) Generalized collinearity diagnostics. J Am Stat Assoc 87:178–183

Freckleton RP (2009) The seven deadly sins of comparative analysis. J Evol Biol 22:1367–1375

Freckleton RP (2011) Dealing with collinearity in behavioural and ecological data: model averaging and the problems of measurement error. Behav Ecol Sociobiol 65:91–101

Freckleton RP, Cooper N, Jetz W (2011) Comparative methods as a statistical fix: the dangers of ignoring an evolutionary model. Am Nat 178:E10–E17

Freckleton RP, Jetz W (2009) Space versus phylogeny: disentangling phylogenetic and spatial signals in comparative data. Proc Roy Soc B—Biol Sci 276:21–30

Garamszegi LZ, Møller AP (2012) Untested assumptions about within-species sample size and missing data in interspecific studies. Behav Ecol Sociobiol 66:1363–1373

Garland T Jr, Ives AR (2000) Using the past to predict the present: confidence intervals for regression equations in phylogenetic comparative methods. Am Nat 155:346–364

Gelman A, Hill J (2007) Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, Cambridge

Grafen A (1989) The phylogenetic regression. Phil Trans Roy Soc Lond B, Biol Sci 326:119–157

Grafen A, Ridley M (1996) Statistical tests for discrete cross-species data. J Theor Biol 183:225–267

Hansen TF (1997) Stabilizing selection and the comparative analysis of adaptation. Evolution 51:1341–1351

Harvey PH, Pagel MD (1991) The comparative method in evolutionary biology. Oxford University Press, Oxford

Ives AR, Garland T Jr (2010) Phylogenetic logistic regression for binary dependent variables. Syst Biol 59:9–26

Martins EP, Diniz-Filho JAF, Housworth EA (2002) Adaptive constraints and the phylogenetic comparative method: a computer simulation test. Evolution 56:1–13

Mundry R (2011) Issues in information theory based statistical inference—a commentary from a frequentist's perspective. Behav Ecol Sociobiol 65:57–68

Nunn CL (2011) The comparative approach in evolutionary anthropology and biology. The University of Chicago Press, Chicago

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

Polly PD, Lawing AM, Fabre A-C, Goswami A (2013) Phylogenetic principal components analysis and geometric morphometrics. Hystrix, Ital J Mammal 24:33–41

Quinn GP, Keough MJ (2002) Experimental designs and data analysis for biologists. Cambridge University Press, Cambridge

Ramsey PH (1980) Exact type 1 error rates for robustness of student's t test with unequal variances. J Educ Stat 5:337–349

R Core Team (2013) R: a language and environment for statistical computing. R foundation for statistical computing. Vienna, Austria

Revell LJ (2009) Size-correction and principal components for interspecific comparative studies. Evolution 63:3258–3268

Revell LJ (2010) Phylogenetic signal and linear regression on species data. Methods Ecol Evol 1:319–329

Rohlf FJ (2006) A comment on phylogenetic correction. Evolution 60:1509–1515

Schielzeth H (2010) Simple means to improve the interpretability of regression coefficients. Meth Ecol Evol 1:103–113

Zuur AF, Ieno EN, Elphick CS (2010) A protocol for data exploration to avoid common statistical problems. Meth Ecol Evol 1:3–14

Part II
Handling Phylogenies in Different Statistical Designs

Chapter 7
Uncertainties Due to Within-Species Variation in Comparative Studies: Measurement Errors and Statistical Weights

László Zsolt Garamszegi

Abstract Comparative studies investigating evolutionary questions are generally concerned with interspecific variation of trait values, while variations observed within species are inherently assumed to be unimportant. However, beside measurement errors, several biological mechanisms (such as behaviors that flexibly change within individuals, differences between sexes or other groups of individuals, spatial, or temporal variations across populations of the same species) can generate considerable variation in the focal characters at the within-species level. Such within-species variations can raise uncertainties and biases in parameter estimates, especially when the data are hierarchically structured along a phylogeny, thus they require appropriate statistical treatment. This chapter reviews different analytical solutions that have been recently developed to account for the unwanted effect of within-species variation. However, I will also emphasize that within-species variation should not necessarily be regarded as a confounder, but in some cases, it can be subject to evolutionary forces and delineate interesting biological questions. The argumentation will be accompanied with a detailed practical material that will help users adopt the methodology to the data at hand.

7.1 Introduction

Since the days of Darwin, questions about evolution have centered around the forces of selection that have led to the extreme diversity in nature we observe today. Current diversity across species is regarded as the result of a large-scale natural experiment, in which their common ancestors had been placed in different environments, and subsequently underwent different selection regimes that

L. Z. Garamszegi (✉)
Department of Evolutionary Ecology, Estación Biológica de Doñana—CSIC,
Av. Américo Vespucio SN, 41092 Sevilla, Spain
e-mail: laszlo.garamszegi@ebd.csic.es

affected their anatomy, physiology, life history, and behavior (Doughty 1996). Therefore, analyzing present-day patterns of interspecific variation enables making inferences about selective mechanisms acting in the past. Accordingly, most comparative studies rely on species as the unit of analysis, and species-specific trait values are subsequently investigated on the branches of the phylogenetic tree. By adopting this focus, most comparative studies inherently assume that species-specific means are biologically meaningful and they can be estimated without error.

In some cases, these assumptions are likely to be met. Take the example of brain size evolution, for instance. Comparing mammals based on species-specific means of relative brain size will probably reveal biologically meaningful comparisons and will allow making inferences about the evolution of cognitive capacities. Every individual within a species has a similar brain size relative to its body size when compared to the level of variation between species. Therefore, obtaining data from Usain Bolt, Albert Einstein, or even from my five-year-old daughter will equally represent the true human-specific value: each will have systematically larger brain-to-body-size ratios than any randomly chosen individual elephant or shrew. However, this assumption becomes less valid if, taking another example, running speed is the trait of interest, which varies more within species. I would bet with high confidence that Usain Bolt could beat an Asian elephant in a race, but I would be extremely worried seeing the same animal running after my daughter. How such variation occurring at the within-species level can affect the phylogenetic findings based on species-specific means?

Historically, the importance of the consideration of within-species variation has been dwarfed by another common assumption: the independence of interspecific data. Species cannot be regarded as independent observations as their shared common ancestry creates varying degrees of similarity between them in their phenotypes; this is classically regarded as the principal confounding factor that can bias interspecific patterns. The phylogenetic association of species, in fact, establishes the essence of comparative studies. Accordingly, a plethora of statistical approaches has been developed to handle such non-independence issues by incorporating the phylogenetic history of species into the analysis. However, the confounding effect of within-species variation remains somewhat neglected, and only recent developments have discovered the significance of this issue in the phylogenetic comparative context.

This chapter aims to bring these statistical developments into the focus of practicing evolutionary biologists. First, I explain what kind of mechanisms can shape variation within species. Next, I show how variation within subjects can alter observed patterns at the between-subject level in both the non-phylogenetic and phylogenetic contexts. Third, I examine how issues about within-species variance and sample size can be dealt with at the different phases of research (research design, diagnostics, core analysis, and interpretation). Following, in addition to the historical development of methods, I review recently proposed phylogenetic models that are able to account for within-species variation. This comparison will highlight the fundamental differences in the theoretical

foundations of these methods, their usefulness for different research designs and data types, and their accessibility in statistical packages. In this section, I will also point about the potential use of within-species variance to address interesting questions rather than as annoying nuisance in analyses. The description of the methods will be accompanied by illustrative biological examples, for which data files and program scripts (mostly in *R*) are made available in the corresponding Online Practical Material (hereafter OPM) at http://www.mpcm-evolution.com. As a closing remark, I discuss how the increasing recognition of within-species variation in phylogenetic comparative studies may lead to a departure from the classical philosophy of focusing on species-specific trait values as being evolutionarily relevant. Such a new direction may also open new horizons for the development of statistical methods.

7.2 Sources of Within-Species Variation

There are several sources that can generate variation, or measurement error, around the species-specific means. Although in a statistical sense, measurement error refers to deviations of any kind that appear between an observed and a true value, such variations can be at least of three types with different biological meaning. Hereafter, I refer to measurement error in a statistical sense meaning all types of within-species variation sensu lato.

First, instrument-related errors or observer effects can cause noise around the mean value of a trait of interest. For example, each equipment or molecular assay has a given precision that delineates a certain confidence range around each measurement. Similarly, estimation outcomes may vary among observers, which also raises an unwanted component of uncertainty when different people assess species-specific means. These deviations cause measurement errors in a narrow sense, since they are unlikely to be associated with the phylogeny and biology of the species at hand, if all species are measured by exactly the same method (or as similar a method as possible). Alternatively, as methods with no doubt vary, different equipment, laboratory assays, and observers may be applied randomly to different species (or at least arbitrarily across species, which is usually the case) to avoid instrument- or observer-related error to raise biases that can affect the underlying biological question. The errors that are introduced by such sources can be assessed by calculating estimates of inter-observer or inter-instrument reliability based on the repeated measure of the same subjects by different observers or via different instrument/assay conditions (e.g., Caro et al. 1979; Reed et al. 2002).

The second type of variation refers to true biological differences at the within-species level. Due to fluctuations in physiology or behavior, individuals of a given species will vary in the expression of certain traits. Even the same individual can demonstrate altering physiological states or produce different behavioral scores in different times or contexts. Moreover, individuals of different age- and/or sex-groups may have particular trait values. Such variation that appears at the within- or

between-individual levels may have biological relevance and can result in non-random fluctuations. For example, higher species-specific means can be associated with higher between-individual variations (e.g., think about body-size variations in mice and elephants), but patterns of within-species distributions may also vary due to several biological reasons and non-independently of phylogeny (e.g., higher variances are preserved in certain closely related taxa). Such non-random patterns make within-species variation due to between- or within-individual variation distinct from instrumental errors and potentially necessitate different treatments. Ideally, if variation at the between-individual level is considerable, several individuals within a species should be measured. This variation can be taken forward into the next levels of analyses (e.g., diagnostics and phylogenetic models), and, unlike in the case of instrument-related errors, should not be assumed as random.

Some type of variation may exist at a higher level. In addition to individuals, populations of the same species can differ and thus can produce deviations around the species-specific means. This is a more challenging problem than the between-individual variation. Different populations may have their own phylogeographical history, and migration between populations can play an additional role in shaping diversity within species, raising challenges in determining the true species-specific trait values (Felsenstein 2002; Ashton 2004). Therefore, to appropriately deal with between-population differences, one may not only need to collect population-specific trait data, but also take into account information about phylogeographical history and gene flow acting at the between-population context (Stone et al. 2011).

Additional complications may appear if the above sources of errors are simultaneously present and generate within-species variances in an additive or more complex manner (i.e., a trait can only be estimated with a given uncertainty due to instrumental errors, but at the same time, it also varies between individuals and populations). Since different types of error have different biological meaning, it might be desirable to treat them separately. Unfortunately, available correction methods handle measurement errors in a broad sense, thus their analytical separation remains currently impractical, and the researcher is left with needing a careful research design and targeted data collection if s/he wishes to discriminate between different types of measurement error in the comparative analysis.

Moreover, when measurement errors exist for more than one trait in the analysis, the potential correlation between these errors can be another confounding issue. For example, if the measurement of two traits relies on the same instrument or observer for a certain group of species, while another set is used for another group of species, there will be an unwanted correlation between measurement errors. Similarly, if between-individual or between-population effects cause similar variations in different traits within species, this will also result in correlated measurement errors. The statistical treatment of correlating measurement errors is achievable in some phylogenetic methods (Ives et al. 2007; Hansen and Bartoszek 2012).

Statistical properties of within-species variation (i.e., how much dispersion from the species-specific trait value exists within species due to within- or between-individual differences of variations between populations) can be

approached by different metrics that describe variations around the mean of a sample. Given that terminologies and abbreviations are used in a somewhat inconsistent manner, I highlight the most relevant definitions in Box 7.1.

Box 7.1 Metrics Describing Within-Species Variation and Sampling Effort

Variance. It is a probability descriptor that measures how far numbers within a sample are spread out. It is measured as the arithmetic mean of the squared differences from the true population mean, thus within-species variance is approached as the average of the squared differences of the within-species data points from the known species-specific trait value:

$$\sigma^2 = \frac{1}{n_i} \sum\nolimits_{i=1}^{n_i} (x_i - \mu)^2$$

where n_i is the within-species sample size (see below), x_i is the individual- or population-specific measure, and μ is the species-specific value. However, given that the true species–species trait value is unknown but is estimated from the available sample of small number of within-species repeats (populations or individuals), the above estimator introduces downward bias. Hence, the following (so-called Bessel's correction) formula should be used to describe within-species variation:

$$s^2 = \frac{1}{n_i - 1} \sum\nolimits_{i=1}^{n_i} (x_i - \bar{x})^2$$

where $\bar{x}$ is the arithmetic mean of the within-species data. Note that σ^2 and s^2 signify variance depending on whether true species-specific values or within-species mean is used for reference. The variance of a variable has units of measurement that are the square of the units of the variable. It is often impractical for interpretation, but most comparative approaches accounting for within-species variation takes data on variation in a form of s^2. For conventional reasons, I present equations based on σ^2 by using the appropriate subscripts to distinguish between within- and between-species variances.

Standard deviation. It is another probability descriptor that measures the dispersion of the distribution, but it is calculated as a square root of variance:

$$\sigma = \sqrt{\frac{1}{n_i} \sum\nolimits_{i=1}^{n_i} (x_i - \mu)^2}$$

for known species-specific values, and

$$s = \sqrt{\frac{1}{n_i - 1} \sum_{i=1}^{n_i} (x_i - \bar{x})^2}$$

for modest within-species samples. Note that although s^2 is an unbiased estimator of variance, s (so-called sample standard deviation) remains a slightly biased estimator for standard deviation. This bias can be considerable for small samples (e.g., $n_i < 10$), but becomes less important at increasing sample sizes. In spite of this, the sample standard deviation is the most commonly used formula, but unbiased sample standard deviation estimators for different distributions are also available.

Unlike for variance, the units of measurement for standard deviation are meaningful on the same scale on which the variable itself was measured. For this reason, interpreting variation within a set of data via standard deviation is more straightforward than via variance. An important aspect of standard deviation is that it is independent of sample size (unlike standard error) and it remains the same at small and large samples.

Measurement (or observational) error. It is simply the difference between a measured value of quantity and its true value. The term has theoretical importance.

Standard error of the mean. It is not a descriptive statistics like standard deviation or variance, but it estimates error bounds on a random sampling process when the true population mean μ is approached by the sample mean $\bar{x}$. By definition, standard error is the standard deviation of the mean of the within-species values and describes how accurately this mean estimate captures the true species-specific value. Note that standard deviation of a sample corresponds to the dispersion of the raw data points around their mean. Another important difference between standard error and standard deviation is that albeit they are measured on the same units (such as the original variable), the former is dependent on sample size. This is because repeated measurements reduces random measurement errors and make the estimator more accurate, thus the mean of the within-species values will be closer to the true species–species estimate as within-species sample size increases. The standard error $\text{SE}_{\bar{x}}$ of the mean can be calculated as:

$$\text{SE}_{\bar{x}} = \frac{s}{\sqrt{n_i}}$$

The method by Ives et al. (2007) uses standard errors. In many cases, the sampling variance is assumed to be the square of the standard errors of the species-specific means (e.g., Hadfield and Nakagawa 2010; Hansen and Bartoszek 2012). However, given that standard errors are sensitive to sample

sizes, correction methods may be applied especially when sample sizes are small and vary among species.

Coefficient of variation. It is a normalized measure of dispersion and describes the spread of data as the standard deviation relative to the sample mean:

$$c_v = \frac{\sigma}{\mu} \quad \text{or} \quad c_v = \frac{s}{\bar{x}}$$

(unbiased estimators for different distributions are available). While variances and standard deviations are interpretable along the scale of the original variable, the coefficient of variation is interpretable independently of the measurement units (i.e., dimensionless) and thus comparable across different traits. This is not only important when comparing different traits, but may be an issue when different species with different species-specific means are contrasted, and larger trait values are accompanied by larger variance. As a general benchmark, distributions with $c_v < 1$ can be interpreted to include low variance, while those with $c_v > 1$ are considered to be loaded with high variance.

Within-species sample size. The number of repeats available within species (i.e., intraspecific sample size), denoted as n_i for species i, and distinguished from N, which is the number of species being compared (i.e., interspecific sample size). n_i can signify the number of populations (or other within-species groups) or the number of individuals that are sampled in a species. Therefore, the total sample size in a comparative study is $\sum_{i=1}^{N} n_i$, which equates with $n_i * N$ if n_i is the same for all species. Within-species sample size is often used as an estimate of sampling effort, because species with large n_i can be considered as species that were studied with higher intensity than species with low n_i, thus they provide more precise estimates for the species-specific trait values. Therefore, within-species sample size can be used to adjust for heterogeneities in sampling effort through the use of statistical weights in the models. Given that $\text{SE}_{\bar{x}} = s/\sqrt{n_i}$, $1/\sqrt{n_i}$ is an important component of the standard error of the mean. For example, if instrumental or between-observer error is constant across species, variation in sample size can still generate differences in measurement error between species. Accordingly, if measurement error data are not available, $1/\sqrt{n_i}$ can be used as an approximation of standard errors (see examples in the text).

7.3 The Statistical Consequences of Ignoring Within-Species Variation

7.3.1 Effects Independent of Phylogeny

7.3.1.1 Increasing statistical noise and attenuation bias

The problem caused by within-subject variation is not unique to phylogenetic comparative methods, but is a well-recognized issue in the statistical literature (Fuller 1987; Bollen 1989; Buonaccorsi 2010). In general, measurement error can introduce uncertainty in the estimates of true values even in very large samples. In a univariate case, imprecise measurements (or true within-subject variation in a broader sense) will increase the error range by which a single measurement approximates the true mean of the sample and will thus decrease our confidence in each datum (Chesher 1991; Manisha 2001). However, such effects act symmetrically on both sides of the distribution, thus measurement error raises random noise but not systematic bias if the focus is to derive estimates for a single variable. This may incur issues about statistical power when, for example, the estimated mean of a sample is contrasted against a hypothetical mean via null-hypothesis testing. Accordingly, in the case of large measurement error, we are more likely to commit type II statistical errors (failing to reject a false null hypothesis) than in the case of small or no measurement errors at the same sample size.

On the other hand, when the relationship between two or more variables is of interest, the presence of measurement error will not only affect precision and statistical significance, but can have considerable influence on parameter estimates, such as correlation or regression slopes (Judge et al. 1985; Fuller 1987; Chesher 1991; Adolph and Hardin 2007). Standard regression and correlation models assume that all variables have been measured correctly, or observed without error (to be more precise, regression models apply these predictions to the predictor but not to the response variables). When particular or all variables have been measured with errors or exhibit variations within subjects, conventional estimates of correlation coefficients (such as Pearson product-moment correlations) will perform with a bias toward zero (i.e., will underestimate true parameters, Fig. 7.1a and b). In a regression problem, such downward bias will be manifested in the underestimation of R^2 and standardized regression coefficients if within-subject variance exists within the response variable, and in the underestimation of unstandardized regression estimates in the presence of error in the predictor variables. The characteristics of this bias are even more complex in nonlinear regressions. The downward bias introduced on parameter estimates by within-subject (e.g., within-species) variability is known as *attenuation bias* and calls for statistical approaches that can account for such a bias (e.g., measurement error regression models, structural equation modeling, and unbiased estimators of correlation coefficient).
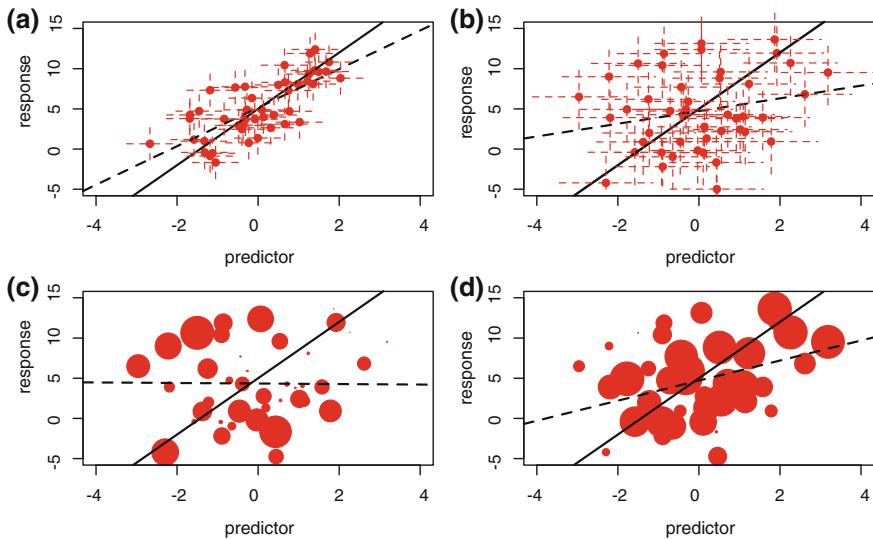
Fig. 7.1 The effect of measurement error and within-species sample size on the estimation of regression slopes from simulations that incorporate different variance components. **a** The within- and between-species variance of both the predictor and response variables are set to provide a repeatability of 0.75 (see Sect 7.4.2.1. about repeatability). **b** The within- and between-species variance of both the predictor and response variables are set to provide a repeatability of 0.40. **c** and **d** same conditions as in **b**, but the data were analyzed by weighted least square regression that relied on two different scenarios for within-species sample sizes. *Dots* represent species, *solid lines* show the regression line (same in all cases) that was used to generate species-specific means, *dashed lines* that are obtained by least square regressions (**a** and **b**: ordinary regression, **c** and **d**: weighted regression) on the simulated data loaded with measurement error, *red crosses* in **a** and **b** define error ranges around both variables that were considered during the simulation, the size of the points in **c** and **d** are proportional to the corresponding within-species sample size

7.3.1.2 Within-Subject and Between-Subject Correlations

The aforementioned situations involve the confounding effects of within-subject variances around the predictor and response that act independently of each other. However, when multiple traits are examined, issues about covariances should also be considered. That is, the apparent relationship between two traits with clustered structure can actually have two components: the between-subject and the within-subject components (Snijders and Bosker 1999, see also Fig. 7.2). The between-subject correlation or regression arises from the relationship between the two variables based on the subject- or group-specific (for us species-specific) mean trait values. The within-subject correlation, however, refers to the associations that appear at the within-subject levels (e.g., from between-individual or -population correlations or regressions). The two components can have opposite direction with effects that contradict each other (for example, in the within-species context,
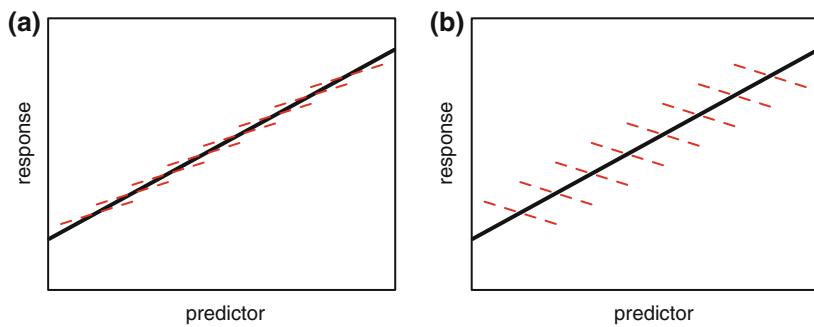
Fig. 7.2 Within-species (*red dashed lines*) and between-species (*black solid lines*) regressions when considering two possible scenarios. **a** both within-species (i.e. as estimated from regressions at the between-individual level) and between-species (i.e. as estimated from regressions using species-specific trait values) regressions are positive, but they somewhat differ in their slopes. **b** within-species regressions are negative while between-species are positive

small-sized individuals live longer than large individuals, while in the between-species context, there is a positive relationship between body size and longevity, such as in Fig. 7.2b). Moreover, the relationships within subjects can vary from subject to subject, which is likely if these represent different species with different ecology. This may be important in a comparative context, if for example, different mechanisms shape between-population patterns in different species, while other selection regimes operate and affect trait associations at the interspecific level. If data are available at the within-subject level, mathematical and statistical solutions are available to separate the within- and between-group components of correlations or regressions (Kreft et al. 1995; Gelman and Hill 2007; Bolker et al. 2009; van de Pol and Wright 2009).

7.3.1.3 Heterogeneity in Sampling Effort

An additional problem met when the unit of analysis (e.g., species) corresponds to repeated measurements of smaller units (individuals, populations) is that due to various constraints on the sampling procedure, different values will often be represented with different within-subject sample sizes (n_i). This will affect another assumption of standard statistical methods, namely that the standard deviation of the error term is constant over all values of the variables (Sokal and Rohlf 1995). This would require each data point to provide equally precise information about the deterministic part of total process variation, a condition that is likely violated if data quality is a function of sampling effort, and subject-specific estimates from small samples will be less reliable than estimates from large samples (Garamszegi and Møller 2010). Such heterogeneity in data quality can be accounted for by using statistical weights in the analyses (generally available for descriptive

statistics of univariate distributions, correlations, and regressions) that give emphasis to each data points according to the underlying sample size (Fig. 7.1c and d).

7.3.2 Effects Evoked by Phylogenetic Associations

The common ancestry of species gives another twist to the story on measurement errors. In addition to the power and attenuation issues detailed above, there are other effects of within-species variation that are called forth only in light of the specific hierarchical structure of the interspecific data. In particular, the above errors interacting with the phylogeny can often enhance the chances of detecting a spurious relationships between variables (Martins 1994; Purvis and Webster 1999; Felsenstein 2004).

7.3.2.1 Mathematical Evidence

Such types of biases are well defined mathematically for the approach based on phylogenetically independent contrasts, when uncontrolled within-species variances are differentially magnified during the standardization of contrasts (Ricklefs and Starck 1996; Felsenstein 2008). The variance of the difference between two species' means depends on the branch length involved (determining phylogenetic variation) and the measurement errors around these means (determining within-species phenotypic variation). By definition, the contrasts are standardized by a quantity proportional to the square root of their total standard deviation (Felsenstein 1985). Such standardization can have a strong influence for closely related species (i.e., when the involved branch lengths are short) if the underlying within-species sample sizes are low (i.e., when measurement errors are large), because the contrasts are divided by far too small quantity. This artificial standardization can produce outlier contrasts for all variables with considerable influence on the estimated regression or correlation coefficients.

7.3.2.2 Simulation Results

Harmon and Losos (2005) presented a simulation study to demonstrate the effect of uncontrolled within-species variation on type I error rates emerging in phylogenetic comparative studies based on independent contrasts. They generated interspecific data for two variables with an expected covariance structure between them for a modest number of species sharing a certain phylogenetic history. They also simulated within-species sampling process around these means by considering different scenarios for sample size in terms of the number of individuals and measurement error. When there is no correlation between two variables, one may

expect to find a statistically significant relationship between them only by chance in 5 % of the simulated datasets. This type I error rate could be reproduced in the non-phylogenetic context, i.e., when species-specific trait values were simulated independently of each other and the units of analysis shared no historical associations. However, the number of significant associations increased above the chance levels in models based on independent contrasts without accounting statistically for within-species variation. At large within-species variation and small sample sizes, type I error rate could increase up to 17 %. The error rates in the phylogenetic simulations could be retained at the chance (5 %) level when most of the variance occurred at the between-species level making intraspecific component negligible and when within-species sampling relied on large sample sizes. Importantly, increasing between-species sample size did not improve error rates, but, in fact, produced spurious correlations at high measurement errors more powerfully than simulations relying on fewer species. Felsenstein (2008) performed similar simulations, in which he used slightly different sampling schemes and found that a true null hypothesis was falsely rejected in about 20 % of the cases when the data were analyzed by the standard contrast method that ignores within-species variance.

If considerable within-species variation is left uncontrolled, parameter estimates become biased when the data are analyzed not only by independent contrasts, but also when the approaches rely on phylogenetic generalized least squares (PGLS). Ives et al. (2007) demonstrated the occurrence of such biases in various evolutionary test situations. For example, they designed a univariate simulation model to study the performance of ancestral state estimators at a given phylogeny and trait value at the root of the tree and with pre-defined within- and between-species variances. They considered an extreme measurement error scenario (being more than two times larger than the standard deviation of the among-species error), which provided evidence that, at a low within-species sample size, the phylogenetic generalized least squares method disregarding within-species variances reveals biased estimates for parameters of evolutionary rate (σ^2) and phylogenetic signal (K). This happens because, by ignoring the within-species component of variation, variability in the data is erroneously attributed solely to the between-species component, which results in the overestimation of the rate parameter and in the underestimation of phylogenetic signal. However, ancestral state estimation seems to be unaffected by the presence of measurement errors.

The influence of within-species variance has also been investigated in bi- or multivariate evolutionary problems. In a correlation model, Ives et al. (2007) showed that an estimator of r without accounting for measurement errors performs with downward bias even when controlling for the phylogenetic association of species. Similar patterns were found in a regression model, in which the established slope parameter was underestimated with a conventional PGLS regression approach when within-species variance was present in the data. Importantly, the degree of this bias was independent of the interspecific sample size, but the confidence intervals around the estimates were narrower when a larger number of

species was involved. Therefore, echoing the results of Harmon and Losos (2005), this indicates that large interspecific sample size when coupled with low intraspecific sample size can accentuate biases with higher statistical power.

7.3.2.3 Empirical Evidence

A meta-analysis of almost two hundred comparative studies investigated empirically if heterogeneity in within-species sample size can have an effect on research findings (Garamszegi and Møller 2010). It appears that in a ~20 % of studies, the use of statistical weights that balance for heterogeneity provides remarkably different results from that of the non-weighted model. This rate is comparable to the effect that a control for phylogenetic non-independence causes. Notably, the result of weighting was particularly strong when there was large variation in sample sizes among species, while homogenization had a minor effect when sample sizes were more balanced among species.

The above studies unanimously suggest that recent issues about within-species variance and sample size should not be taken as a false alarm, as there are situations when the variability of trait values within species can have a considerable effect on the estimation of parameters with evolutionary importance. However, most of the simulations were made under conditions in which extreme error structures were created: i.e., within-species samples consisting of few individuals and variances that are comparable with the between-species variances. Most of the comparative studies in practice might meet with more relaxed conditions when conventional phylogenetic methods can also perform with tolerable type I error rate even by ignoring within-species components of variations. This does not mean that the problem can be ignored, but highlights the importance of the consideration of the measurement error issue at the level of study design and data diagnostics, and its appropriate treatment at the analytical level if the data at hand require. Such an empirical approach to within-species variance resembles the philosophy that is similar to how we nowadays handle the confounding effects due to common descent: we estimate the phylogenetic signal in the data first (more precisely in the model residuals, see Chap. 5) and then obtain parameter estimates from the model at the estimated signal value (Freckleton 2009).

7.3.3 Phylogenetic and Other Biological Confounders

So far, within-species variances and sample sizes were treated as if they were fluctuating at random with respect to the biological question under testing. However, there might be cases when properties of between-individual or between-population distributions are shaped by evolutionary forces; thus, within-species

sample sizes and/or variances are not necessarily independent of the phylogenetic history of species or other biological predictors. For example, Garamszegi and Møller (2011) claimed that within-species sample size (also contributing to measurement errors via $1/\sqrt{n_i}$, see Box 7.1) can vary along several species–species characteristics that determines the probability of sampling of individuals. This is because some species may occur at low abundance, display trap-shy behaviors or have a life history that makes their sampling difficult, which would all decrease their probabilities of being sampled. If such probabilities are determined by certain phylogenetic or biological attributes of species, this will render within-species sample sizes to be related to the same attributes. From the empirical part of the argument (Garamszegi and Møller 2012), it was evident that sampling effort in terms of within-species sample size was consistent across different studies on birds using different sampling methods suggesting that available sample sizes are species-specific attributes. Moreover, it was dependent on abundance, body mass, and predator avoidance behavior implying that applicable sampling effort is determined by some biological properties of the species.

7.4 The Statistical Treatment of Within-Species Variation in Phylogenetic Comparative Studies

The problems posed by within-species sampling variance can be considered in each phase of research, from the design of studies to the interpretation of the results. These steps will be discussed below.

7.4.1 Study Design: Balancing Sample Sizes

When designing a phylogenetic comparative study, the observer is usually restricted to balance between the within-species and the between-species sample sizes (Harmon and Losos 2005). A certain number of species is important to reach a sufficient statistical power in the analyses, but also to make generalizable evolutionary inferences. On the other hand, simulations (Harmon and Losos 2005; Ives et al. 2007) show that low within-species sample size can bias parameter estimates even when several species are included. Thus, sample size requirements at different levels to deal with power and bias are often in conflict with each other. Therefore, some efforts should be devoted to the appropriate within-species sampling, event at the cost of decreasing between-species sample size. Ideally, these constraints can be optimized in a pilot study, in which the variances at the between- and within-species levels can be determined in a subsample of species. Such information can subsequently guide the investigator when determining sample sizes, for which the simulation results might serve some rules of thumb (Harmon and Losos 2005; Ives et al. 2007; Felsenstein 2008). In general, when within-species variance appears

negligible compared to the among-species variance in a pilot study, it may be a convincing evidence for that sampling effort at the within-species level will be less important on a wider scale. Hence, the researcher is safe to conclude that a few estimates per species provide a reliable representation of the species-specific mean trait values. However, when a considerable proportion of variance is at the within-species level (see quantitative estimation in the next section), multiple measurements are warranted for each species, and the within-species variance should be taken forward into the comparative analyses.

Due to various biological reasons, some species are easier to sample than others. Therefore, it is unrealistic to obtain the same within-species sample size in every species, which sets up additional challenges for study design. Shall we devote more research effort to sample less accessible species at the cost of lowering the sample size for a large number of more common species? Finding an answer to such questions can be a complex task and may require considerations about the biological problem and the phylogeny at hand. Some pilot studies as well as phylogenetic targeting (see Arnold and Nunn 2010) may also be of help to optimize research effort.

7.4.2 Diagnostics

7.4.2.1 Repeatability

For a comparative study of species-specific means to be meaningful, these means are required not to be confounded by measurement errors, and if this assumption is fulfilled, within-species variance could be omitted in the analyses. Although, most comparative analyses apply this omission, the potentially confounding effect of measurement error is rarely considered in practice (and even in theory). If species-specific values cannot be estimated without errors, within-species variances (or other components of errors) need to be incorporated in the phylogenetic models. In this case, the researcher needs to invest a substantial effort in collecting a sufficiently large sample for each species that would allow to capture trait variations therein and to obtain good estimates for species means (or other statistics). To make decisions about such investments and about the subsequent analytical strategy, it might be useful to have a glance about different components of the variance prior to the phylogenetic analyses. If based on the measurement of different individuals or populations, multiple data are available for, at least, a subset of species (e.g., from a pilot study), the ratio between the within-species and between-species components of variances can be assessed by calculating repeatability.

Repeatability is the intraclass correlation coefficient that can be derived from different variance components, mathematically as the proportion of the between-subject variance relative to the total variance (Sokal and Rohlf 1995):

$$R = \frac{\sigma_{\text{between}}^2}{\sigma_{\text{between}}^2 + \sigma_{\text{within}}^2}, \tag{7.1}$$

where

$\sigma_{\text{between}}^2$ between-species variance and
σ_{within}^2 within-species variance.

If most of the variance resides at the between-species level and within-species variance remains minuscule, the ratio will approach unity. However, if within-species variation is similar to the between-species variation, the estimator will return values that are ~0.5 and will approximate zero if within-species variation is of a major concern. Repeatability can offer a standard metric to help the observer judge about how much s/he can trust in the species-specific trait estimate as a meaningful unit for the comparative analysis. Working with a trait that has repeatability close to one indicates that a single individual will well-reflect the true species-specific value. On the other hand, low repeatability indicates that within-species variance may warrant some attention, and for an appropriate statistical treatment, a systematic within-species sampling is needed to capture such variation. Behavioral or physiological traits may often involve such a modest within-species repeatability, while morphological or life history characters may depict less variation within species incurring higher repeatability (see also Blomberg et al. 2003).

Classically, repeatability can be acquired from an ANOVA model, in which the variation in the response measurements is partitioned into components that correspond to different sources of variation (Lessells and Boag 1987). The widely used ANOVA-based repeatability takes the mean between-subject sum of squares and the mean within-subject (residual) sum of squares and also considers the species-specific sample sizes. Recently, Nakagawa and Schielzeth (2010) provided a list of functions for repeatability (together with its confidence interval, and a test statistics contrasting the estimated value against zero) based on mixed model approach that can be flexibly used for a variety of data types and distributions. These approaches can also be applied in the comparative context, where the interest is to determine whether species consistently differ in their mean trait value (some strategies are given in the OPM). However, the phylogenetic relationships of species may warrant some attention and potentially necessitate the consideration of more complex hierarchical modelling.

7.4.2.2 Independence

Another diagnostics that may be useful before entering into the core phylogenetic analyses is to assess whether within-species sample size (or variance) occurs randomly. There might be several biological reasons to why some species are systematically easier to sample resulting in larger sample sizes than other species (Garamszegi and Møller 2011). Such factors can be associated with species-specific

abundance (rare vs. common species), behavior (species that avoid traps vs. species that are attracted by them), as well as life history and ecology (species breeding in accessible vs. inaccessible habitats, solitary vs. colonial species) and even morphology (small vs. large species with different mobility and/or detectability). Furthermore, these effects may vary in non-random manner with respect to the phylogenetic associations of species. Hence, within-species sample sizes can depict a phylogenetic structure at the across-species level. These deviations from randomness can be investigated by testing if within-species sample size (or variation) is interspecifically related to biological predictors and phylogeny. (see Garamszegi and Møller (2012) for an example analysis)

7.4.2.3 Preparation of Data

Before testing the evolutionary hypotheses in the statistical models, some preparatory steps may be warranted. These may include the conventional exercises for data diagnostics and the verification of model assumptions (e.g., distribution, collinearity, balanced design, see more in Chap. 6), but also efforts to make our data suitable for the particular method that will be used to control for within-species variance. For example, we may need to decide whether we aim to work with individual (population)- or species-specific data. We can use data on individuals for the contrast and likelihood surface methods (see Table 7.1 and text below), but the calculation of repeatability discussed above also requires repeated measurement within each species and the underlying dataset needs to be tabulated accordingly. We can supply species-specific trait values for the other methods such as based on PGLS regression techniques (see Table 7.1 and text below). Means at the species level can be either calculated from the raw individual data through simple summary statistics or should already be available in this form if obtained from other sources.

When working with species-specific datasets, most of the methods assume that within-species variances are known and correspond to large samples. However, this criterion might be violated in most of the comparative studies, in which within-species samples are often limited to few individuals (or populations). It is also common that sample size at this level equals one, which makes variances mathematically inestimable. For such a case, a corrected estimate might be desired. For this purpose, Ives et al. (2007) suggest first calculating a pooled variance over the entire sample:

$$\bar{\sigma}^2 = \frac{1}{N_{\text{species}}} \sum_{i=1}^{N_{\text{species}}} \sigma^2_{\text{within}_i}, \tag{7.2}$$

where

$\bar{\sigma}^2$ weighted (pooled) within-species variance,

$\sigma^2_{\text{within}_i}$ within-species variance observed in ith species, and

N_{species} interspecific sample size after excluding species with a single observation.

Hansen and Bartoszek (2012) suggested an improved estimate of the pooled variance by weighing the species with their sample sizes:

$$\bar{\sigma}^2 = \frac{\sum_{i=1}^{N_{\text{species}}} \sigma_{\text{within}_i}^2 (n_i - 1)}{\sum_{i=1}^{N_{\text{species}}} (n_i - 1)}, \tag{7.3}$$

where

n_i within-species sample size for the ith species, other abbreviations as in Eq. (7.2).

With these estimates, the species-specific standard errors (SE_{n_i}) as well as variances could be computed:

$$\text{SE}_{n_i} = \frac{\bar{\sigma}}{\sqrt{n_i}} \tag{7.4}$$

If the model for comparative analysis requires species-specific variances ($\bar{\sigma}_{n_i}^2$) to be entered, these could be approximated as the square of this standard error ($\text{SE}_{n_i}^2$), that gives:

$$\bar{\sigma}_{n_i}^2 = \frac{\bar{\sigma}^2}{n_i}, \tag{7.5}$$

The use of this replacement procedure might be problematic, if true variances vary considerably among species due to scaling effects for example (e.g., consider body mass variation in a mouse and an elephant species). In this case, unrealistically high variances would be assigned to species, which, in reality, could actually be characterized by small variance (e.g., variance in mouse would be adjusted partially based on variance in elephant). This problem can be reduced by applying a log-transformation on variances (or other variance-stabilizing transformation) before the adjustment (see below) or by using the scale-independent coefficient variation (cv) parameter for the pooling and subsequent weighting procedure, from which error variances could be back-calculated (see Box 7.1 for calculations).

For making meaningful interpretations from interspecific patterns, tip values are often required to have a scale on a log-axis for both statistical and biological reasons. In such a case, not only species-specific means, but also the associated variances should also be transformed. Note that this transformation also normalizes within-species variation, thus diminishes the problems caused by unequal variances due to scaling effects (Revell 2010), as discussed above. Log-normal distributions require that the joint log-transformation of the mean and variance occurs according to the following approximations equations:

$$y_i = \log\left(\frac{\bar{x}_i^2}{\sqrt{\sigma_{\text{within}_i}^2 + \bar{x}_i^2}}\right) \tag{7.6a}$$

$$v_i = \log\left(1 + \frac{\sigma_{\text{within}_i}^2}{\bar{x}_i^2}\right), \tag{7.6b}$$

where

$\bar{x}_i$ and $\sigma_{\text{within}_i}^2$ mean and variance, respectively, on the original scale,

y_i and v_i mean and variance, respectively, on the log-scale for species i.

However, these approximations are not necessary if the individual level data are accessible, when one can just compute the mean and variance on the log-scale directly.

In cases when we are interested in assessing the effect of the heterogeneity in species-specific sample sizes, it might be important to consider issues about the transformation of within-species sample sizes (n_i). Data heterogeneity is presumably more influential in cases of low sample sizes. On the other hand, after a certain level, increasing sample size has a minor effect on precision. Therefore, we may want to downweight data points with very low sample sizes, but without making too much discrimination between species-specific trait estimates that come from a reasonably large within-species samples (Garamszegi and Møller 2010). Accordingly, a logarithmic or square-root transformation on sample sizes may help dealing with this problem (see also Chap. 12).

Examples for calculating measurement errors based on pooled variances and for log-transformation are given in the OPM.

7.4.3 Incorporating Within-Species Variation into the Phylogenetic Analysis of Species-Specific Traits

If we cannot achieve negligible within-species variances through a careful study design, and a repeatability analysis indicates that the measurement error on any of the investigated traits is considerable and may be meaningful, we need to use phylogenetic comparative studies that can account for such variance components. Assuming that information on the dispersion of data at the within-species level is available in any form (e.g., as raw individual data or as a probability descriptor summarized in Box 7.1), different phylogenetic methods can be applied to deal with different test situations and data types with each offering different benefits. In the sections below, I will provide an overview on these methods. I start this

revision from the historical perspective, as I find it important to get a picture about the essence of the classical methods in order to understand how modern approaches discussed subsequently work.

7.4.3.1 The History of Interspecific Comparative Methods That Account for Within-Species Variation: Back to the Root

The Autoregression Model

Although the spread of comparative methods with measurement error can be witnessed in the contemporary literature, the history of the underlying methodology goes back to the beginnings of phylogenetic comparative methods. In the same year when Felsenstein's (1985) seminal paper was published, Cheverud et al. (1985) proposed an alternative approach to control for phylogenetic non-independence. This method was based on an autoregression model to accommodate the concept of phylogenetic constraints in interspecific studies through partitioning the observed total trait variance into inherited (phylogenetically determined) and taxon-specific (caused by independent evolution) components:

$$y_i = p\mathbf{W}y_i + e_i, \tag{7.7}$$

where
y_i observed trait mean for species i taken from the $\mathbf{y}$ vector of standardized trait values for N_{species} species,
p phylogenetic autocorrelation coefficient (scalar),
$\mathbf{W}$ phylogenetic connectivity matrix reflecting the relatedness of species (i.e., genetic correlation between species),
e_i ith element of the $\mathbf{e}$ vector of residuals.

In this equation, $p\mathbf{W}y_i$ represent the phylogenetic part, while e_i stands for the taxon-specific part. Originally, the autocorrelation model does not require the specification of an evolutionary model (such as Brownian motion), it only relies a relaxed assumption that the inherited component is similar among closely related species and different among distant species (but in principle, different evolutionary assumptions could be brought into the $\mathbf{W}$ matrix). The approach by Cheverud and coworkers (1985) includes estimation procedures based on maximum likelihood (ML) iteration for the phylogenetic autocorrelation parameter p, which can be used to make inferences about the importance of phylogeny in trait evolution. This methodology has not received as much popularity in practice as the independent contrast method (Felsenstein 1985), but its inherent promise for incorporating issues about within-species variation has been recognized in its subsequent analytical development toward greater flexibility (e.g., Gittleman and Kot 1990). Along this line, Cornillon et al. (2000) considered issues about differences between

populations and, accordingly, split the inherited variance part into inter- and the intraspecific components, while they also expressed the residual part at the level of population (and not species). These matrices were used to fit an autoregressive model through a maximum likelihood (ML). By doing so, the method is able to capture within-specific variation that occurs among populations in both univariate and multivariate test situations.

The Phylogenetic Mixed Model of Lynch and Its Extensions

The main logic of Cheverud et al. (1985) was also influential on the improvement of comparative methodologies of other types. Lynch (1991) extended mixed modeling techniques taken from quantitative genetics to decompose observed mean pheno-types into different components. His model was based on the general formula:

$$y_i = \beta_0 + a_i + e_i \tag{7.8}$$

$$\mathbf{a} \sim \mathcal{N}\left(0, \sigma^2 \mathbf{C}\right) \tag{7.8a}$$

$$\mathbf{e} \sim \mathcal{N}\left(0, \sigma_e^2 \mathbf{I}\right), \tag{7.8b}$$

where
y_i observed trait mean for species i taken from the $N_{\text{species}} \times 1$ dimensioned $\mathbf{y}$ vector of standardized trait values,
β_0 grand mean of the character over the whole phylogeny (intercept),
a_i ith element of the $\mathbf{a}$ vector of heritable additive values with a dimension of $N_{\text{species}} \times 1$,
e_i ith element of the $\mathbf{e}$ vector of residuals (dimension: $N_{\text{species}} \times 1$),
$\mathcal{N}$ signifies that values of the given vector (i.e. $\mathbf{a}$) are taken from normal distribution that is specified with a mean (i.e. 0) and variance (i.e. $0, \sigma^2\mathbf{C}$)
σ^2 overall phylogenetically inherited variance (rate of evolution),
σ_e^2 residual variance,
$\mathbf{I}$ identity matrix (dimension: $N_{\text{species}} \times N_{\text{species}}$),
$\mathbf{C}$ correlation structure defined by the phylogeny (dimension: $N_{\text{species}} \times N_{\text{species}}$).

In this approach, a_i represents the heritable phylogenetic effect sensu Cheverud et al. (1985), while e_i is Cheverud's species-specific effect that also includes sampling error beside the nonadditive genetic effects and environmental effects. Lynch suggested an iterative approach based on expectation-maximization (EM, Dempster et al. 1977) algorithm to find models with maximum likelihood (ML) that can be used to estimate parameters, such as the mean phenotypes of ancestral taxa, additive values, and residuals deviations as well as the variance-covariance structure of the components of taxon-specific means. These parameters can serve basis for computing regression coefficients and hypothesis testing. Although, in the above model the author generally assumed that within-species variation is negligible and

sampling errors on the phenotypic means can be treated as zero, he pointed that such an assumption may be violated in many cases, for which the original method could be adjusted if data on sampling variances and covariances are available.

Such a premise was further exploited by Christman et al. (1997) and Housworth et al. (2004). These authors argued that if the estimation of species-specific character states from a sample of individuals or populations is subject to considerable error due to the relatively small number of individuals sampled per species, an additional component should be added to Lynch's formula to factor out the within-species variances. The model of Housworth et al. (2004) for univariate case yields

$$y_{ij} = \beta_0 + a_i + e_i + \varepsilon_{ij} \tag{7.9}$$

$$\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\sigma}_\varepsilon^2 \mathbf{I}), \tag{7.9a}$$

where

y_{ij}	observed trait value for individual j in species i
ε_{ij}	individual error term that is associated with the measurement of individual j in species i,
$\boldsymbol{\sigma}_\varepsilon^2$	variance caused by errors when measuring a single individual, the corresponding $\mathbf{I}$ identity matrix has a dimension of dimension: $\sum_{i=1}^{N_{\text{species}}} n_i \times \sum_{i=1}^{N_{\text{species}}} n_i$,
β_0, a_i, and e_i	intercept, phylogenetic, and non-heritable residual components, respectively, as in Eq. (7.8).

Christman et al. (1997) present an illustrative analysis (and the corresponding *MATLAB* codes) on morphological characters originating from four populations of amphipods, in which they relied on the extended Lynch's model but without the non-heritable effects ($y_{ij} = \beta_0 + a_i + \varepsilon_{ij}$). Through the incorporation of the term ε_{ij}, these approaches bring the focus onto individuals as the unit of analysis assuming that each species on the phylogenetic tree is formed by a hard polytomy of individuals. The length of the within-species branches scales with the degree of measurement error and needs to be estimated in parallel with other parameters in the model.

Regression Techniques Based on Phylogenetic Generalized Least Squares (PGLS)

The autoregressive method of Cheverud et al. (1985) and the mixed model approach originating from Lynch (1991) as well as their derivates combine the phylogenetic constraint with the statistical model via a mean structure in the equation ($p\mathbf{W}y$ or a_i), while the intraspecific variance is usually lumped within the error term (Lynch 1991; Christman et al. 1997; Housworth et al. 2004). As an alternative approach, phylogenetic methods based on generalized least squares (GLS) models incorporate the phylogeny through the error structure (Grafen 1989; Martins and Hansen 1997;

Revell 2010, see Chap. 5). This solution offers a flexible combination of errors from different sources (e.g., phylogeny and within-species variance) as well as to accommodate various test situations (e.g., estimating ancestral states, rates of evolution, phylogenetic effects, correlations, and regressions). Martins and Hansen (1997) present a general linear model in the form of:

$$\mathbf{y} = \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\varepsilon}, \tag{7.10}$$

where

$\mathbf{y}$ vector of characters or functions of character states for extant or ancestral taxa,

$\boldsymbol{\beta}$ vector of regression slopes,

$\mathbf{X}$ matrix of states of other characters, environmental variables, phylogenetic distances, or a combination of these, and

$\boldsymbol{\varepsilon}$ vector describing error structure due to various sources.

This equation can be broadly used to translate evolutionary questions into a statistical formula. By a smart definition of $\mathbf{X}$ and $\mathbf{y}$, several evolutionary problems can be tackled, while the characterization of $\boldsymbol{\varepsilon}$ allows describing the impact of confounding factors that can cause noises or biases in the estimation of $\boldsymbol{\beta}$. The error structure is composed of at least three types of error: the error due to common ancestry ($\boldsymbol{\varepsilon}_S$), the error due to within-species variation (on any character, $\boldsymbol{\varepsilon}_M$), and error due to the uncertainty in the reconstruction of phylogenetic history ($\boldsymbol{\varepsilon}_P$). These errors can be combined in the PGLS framework (see Chap. 5), which thus allows the careful definition of residuals that accommodate a complex covariance structure within the term ε.

Practical Constraints

Despite the relatively well-established statistical background for treating within-species variance in different comparative approaches (e.g., autoregressive models, phylogenetic mixed models borrowed from quantitative genetics, and phylogenetic generalized linear models), until recently, these statistical approaches were rarely exploited in practice. The reasons for such ignorance may rely on the practical intractability of the proposed algorithms, the lack of evidence pointing to the confounding role of measurement errors in the interspecific context and the scarcity of data that captures within-species patterns. For example, the expectation-maximization (EM, Dempster et al. 1977) method proposed by Lynch (1991) to fit models was very slow in practice, and even the reparameterized algorithm that remedies this problem can only be applied to uni- and bivariate cases. The first widely accessible computer software for incorporating measurement error was *Compare* (Martins 2004). When it became accessible to deal with intraspecific variance, the phylogenetic independent contrast method (that neglects such variance) was already flourishing in its renaissance epoch thanks to easy access to the

program CAIC. (Purvis and Rambaut 1995). In addition to these technical challenges, early practitioners of the comparative approach might have disregarded the importance of within-species variation because its confounding effects remained under-documented compared to the well-known biases that can be caused by common descent. Finally, various constraints during the collection of interspecific data and the assembly of phylogeny may have shifted the focus from the within- to the between-species patterns preventing the adoption of measurement error models by appropriate within-species data.

Nonetheless, even in this early phase of the history, some investigators did consider issues about measurement error in their comparative study, and their solutions deserve mentioning. In a comparative study on the relationship between population density and body size in birds, Taper and Marquet (1996) corrected estimated regression parameters for attenuation bias due to measurement error based on the ratio of the total variance and between-subject variance (Madansky 1959) and concluded that such a correction had a minimal effect on the focal relationship. However, this correction is only applicable to the parameter estimates of the ordinary least square regressions, thus the authors could only employ it in the analysis based on raw species data without adjusting for phylogeny (but see its extensions below for PGLS sensu Hansen and Bartoszek 2012). In another pioneering study, Monkkonen and Martin (2000) relied on randomization and bootstrapping procedures to investigate the influence of the between-population trait variations on the interspecific relationship between clutch size and nest excavation propensity in *Parus* tits. They found that the outcome of the analysis was largely similar across the 1,000 bootstrapped samples that randomly picked one population estimate for each trait for each species. Although such a resampling technique appears to be able to incorporate uncertainty in parameter estimates, it may not be useful as a general method to control for within-species variance, because it is not able to cope with attenuation bias (i.e., correct for the degree of underestimation of parameters). When intraspecific variance is considerable, the confidence range around the regression slope is more severely biased than the regression slope that is based on species-specific mean values (Fig. 7.3).

7.4.3.2 Recently Developed Comparative Methodology for Handling Within-Species Variance Components: Getting into Practice

Modern phylogenetic approaches began to recognize the importance of the statistical problems that can be caused by measurement errors, and such considerations gave a burst to the expansion of available softwares and also enhanced the spread of the methodology into research practice. These approaches, while taking into account within-species variance, can now accommodate a wide range of evolutionary questions about correlated trait evolution, ancestral states and phylogenetic signals (although the corresponding methodologies are not equally developed). Most of these new approaches are closely linked with classical

Table 7.1 Recently proposed phylogenetic comparative methods that can account for intra-specific variances due to measurement error or biological variations at the within- and between-individual as well as the between-population level

Approach	Reference	Test situation	Data type	Software
Independent Contrasts	Felsenstein 2008	Multivariate: Phylogenetic correlation Regression	Raw individual-specific data	Phylip R packages: *varCompPhylip (ape)* *pic.ortho (ape)*
PGLS	Ives et al. 2007 Martins & Hansen 1997	Univariate: Ancestral state estimation Rates of evolution Phylogenetic signal Multivariate: Phylogenetic correlation Regression Reduced major axis regression	Species-specific data on within-species variance or standard error Taxon-wise estimate on the universally applicable within-species variance or standard error Taxon-wise estimate on the universally applicable within-species variance or standard error that is weighted by 1/sample size Raw individual-specific data to calculate the above variance/error components	Matlab codes from authors Compare 4.6 R packages: *gls (nlme & caper)* *pgls.Ives (phytools)* *phylosig (phytools)* *fitContinuous (geiger)*
PGLS	Hansen and Bartoszek 2012	Multivariate: Regression	Species-specific data on within-species variance or standard error	R packages: *Slouch* (from author) *GLSME* (from author)
Likelihood computation (simulation-based)	Kutsukake and Innan 2012	Univariate: Ancestral state estimation Rates of evolution Estimation of evolutionary model parameters Multivariate: (could be extended)	Species-specific data on within-species variance or standard deviation Alternatives to normal distribution can be accommodated	a program written in C language (from author)
Likelihood computation (Bayesian)	Revell and Reynolds 2012	Univariate: Ancestral state estimation Rates of evolution Estimation of evolutionary model parameters	Raw individual-specific data	R packages: *fitBayes (phytools)*
Mixed models	Hadfield and Nakagawa 2010	Multivariate: Regression	Species-specific data on within-species variance or standard error	R packages: *MCMCglmm* (MCMCglmm)

methods and can be grouped into the following main (not necessarily exclusive) categories, which are also listed in Table 7.1. The performance of these methods on simulated data is demonstrated on Fig. 7.4. Worked examples can be found in the OPM.

Independent Contrasts

A new method based on independent contrasts (Felsenstein 2008) can be considered as a modified version of the Lynch's model (Lynch 1991) under the assumption that the evolution of species-specific values depicts Brownian motion with a perfect phylogenetic heritability component. This extended contrast method allows for multiple individual measurements per species resulting in within-species phenotypic variances that are greater than zero and are the same for all species. If phenotypic values are available for individuals, the contrasts can be computed at both the within- and the between-species levels. This computation assumes that individuals within a species are connected to each other with zero branch lengths to form a species-specific node on the phylogenetic tree.
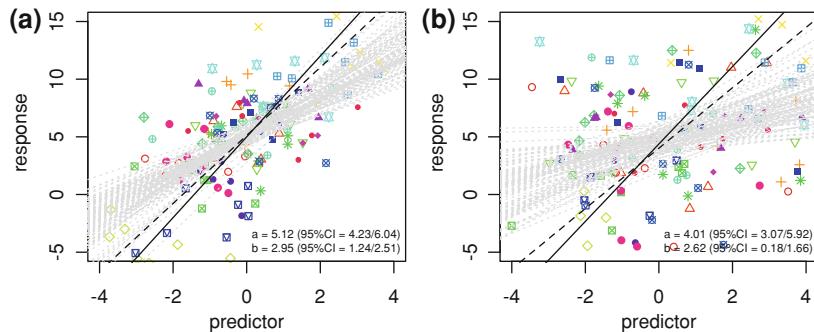
Fig. 7.3 Dealing with within-species variation in interspecific comparisons by resampling technique. Each data point represents an individual within species that are coded by different symbols. The interspecific relationship between the predictor and response is estimated by randomly taking one individual from each species to represent the species-specific character values, which are then regressed across species. This procedure can be repeated multiple times (100–1000). The resulting regression lines are given by the *gray dotted lines* (from 100 resamplings). The *dashed black line* shows the regression line that could be obtained by using the averaged individual trait values at the species level. The slope and intercept of this line are given in the legend together with the confidence intervals that could be estimated from the resampling of individual values. *Solid black line* shows the true regression line (a = 5 and b = 3.5) that originates from the generating species–species values, around which the individual-specific data were simulated under two different variance scenarios: **a** small within-species variance, **b** large within-species variance

Given such a tree structure linking all individuals of all species, the contrasts can be obtained in the classical way (e.g., Felsenstein 1985) with the exception that they are not standardized by their variances but are multiplied by coefficients that include a weight factor for the number of observations. Specifically, the modified method does not scale the contrasts to have equal variances, but it rather applies an orthonormal transformation on the original variables so that the sum of squares of the coefficients in the contrasts is forced to add up to one. Under such constraints, at the within-species level, contrasts can be written as (Felsenstein 1985; Paradis 2011):

$$
\begin{aligned}
c_{i_1} &= \sqrt{\frac{1}{2}}(y_{i_1} - y_{i_2}) \\
c_{i_2} &= \sqrt{\frac{2}{3}}\left(y_{i_3} - \frac{y_{i_1} + y_{i_2}}{2}\right) \\
&\;\vdots \\
c_{i_{n_i-1}} &= \sqrt{\frac{n_i - 1}{n_i}}\left(y_{i_{n_i}} - \frac{1}{n_i - 1}\sum_{j=1}^{n_i-1} y_{ij}\right)
\end{aligned}
\tag{7.11}
$$
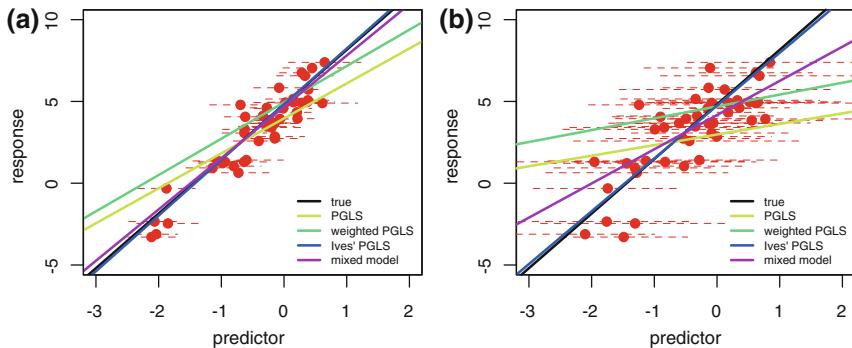
Fig. 7.4 Dealing with within-species variation in interspecific comparisons by modern phylogenetic methods (listed on Table 7.1). Each data point represents a species-specific mean that that can be measured with a certain precision as estimated by the within-species variance on the predictor (for simplicity, only errors on the predictors were considered). The interspecific relationship between the predictor and response was assessed by (1) Phylogenetic least squares methods without considering within-species variance (*yellow line*), (2) weighted phylogenetic least squares models that give emphasis on different data points according to the underlying variance (*green lines*), (3) measurement error PGLS models as was suggested by Ives et al. (2007) (*blue line*), and (4) phylogenetic mixed models (*purple line*). Solid black line shows the true regression line (a = 5 and b = 3.5) that originates from the generating species–species values, around which individual-specific data were simulated under two different variance scenarios: **a** small within-species variance (repeatability ~0.75) and **b** large within-species variance (repeatability ~0.25) on the predictor. *Dashed horizontal lines* indicate the degree of intraspecific variation within the predictor

where

c_i	$1 \ldots n_{i-1}$ within-species contrasts for species i,
n_i	number of individuals measured in species i,
y_i	$1 \ldots n_i$ individual observations for species i, and
$\sqrt{\frac{n_i-1}{n_i}}$	general form of the normalizing constant.

Accordingly, having four observations in a species for example, three contrasts can be derived with normalizing constants $\sqrt{1/2}$, $\sqrt{2/3}$, and $\sqrt{3/4}$, respectively. Considering alternative branching patterns within species, there can be several ways to calculate the orthonormal contrasts, each resulting in the linear combination of the original measurements taken at the individuals that defines the species-specific character states at the end. Taking the orthogonality constraints into account, Felsenstein further developed an algorithm for computing between-species contrasts recursively. In this approach, for each character, within- and between-species sets of contrasts can be derived fulfilling the constraint that different contrasts for same characters are independent. Therefore, the same contrasts for different characters will have the covariance that is equal to the covariance of the original character values.

Unlike the contrasts that are obtained by the original method relying on species means, the contrasts derived from individual data cannot directly be imported to conventional statistical approaches to estimate parameters of evolutionary importance (e.g., a slope from a regression forced through the origin). This is because covariances are composed of different components at the within- and between-species level (i.e., phenotypic covariance arising from between-individual associations and covariance due to convergent evolution, Fig. 7.2). However, the new contrast method also includes an expectation-maximization (EM) algorithm (Dempster et al. 1977) to partition observed covariances among traits, similarly to Lynch's model, into phylogenetic (i.e., the between-species covariances) and a phenotypic (i.e., the within-species covariances) components according to the model

$$\mathbf{T} \otimes \mathbf{A} + \mathbf{I} \otimes \mathrm{P}, \tag{7.12}$$

where

$\mathbf{T}$ phylogeny matrix (expected covariances based on the length of shared branches),

$\mathbf{A}$ between-species (phylogenetic) covariance matrix,

$\mathbf{P}$ within-species (phenotypic) covariance matrix,

$\mathbf{I}$ an identity matrix,

$\otimes$ Kronecker product multiplication (each element of the first matrix is multiplied by each element of the second matrix).

The elements of the $\mathbf{A}$ and $\mathbf{P}$ matrices can be estimated from the contrasts (see for example in the OPM relying on *varCompPhylip* that calls functions from program *Phylip*) and can subsequently be used for making inferences about the coevolution of traits. For example, if the aim is to challenge the null hypothesis that species-specific values of two (or more) traits vary independently of each other with regard to the phylogeny while accounting for the potential within-species covariation of traits, one can fit a model in which the elements of $\mathbf{A}_0$ are forced to be zero. This model can be compared by using likelihood ratio test with the model that is based on an $\mathbf{A}_1$ matrix that represents the true associations between species due to common ancestry. Estimated covariances from the model with the highest likelihood can also be used to obtain parameters from regressions of the variables on each other or correlation coefficients that are not confounded by phylogenetic associations and within-species covariances.

Extensions to the PGLS

Approaches based on phylogenetic generalized least squares provide a rich set of tools to study the effect of measurement error in phylogenetic comparative studies. In these models, following the logic of Martins and Hansen (1997), within-species

variance is lumped within an error term in the regression equation. Therefore, individual-specific estimates are not required, neither do assumptions about phylogenetic resolutions within species. The model can flexibly accommodate information on within-species variation. The flexibility is also prevalent in the fact that the model allows measurement errors to be same or different for different taxa, and actually, it assumes that within-species variances are available without bias and not needed to be estimated. Of further advantage is that the PGLS approach can not only applied to investigate questions about the correlated evolution of traits, but can also be tailored to various test situations.

The PGLS approach for accounting measurement error in interspecific comparative studies relying on species as the unit of analysis adheres to the following logic. Using information on the phylogeny and within-species variances (both are known from data), and considerations about how to combine different error components due to phylogenetic effects and within-species variances (and other sources), one can establish an overall variance structure to describe the expected covariance matrix in the models' residual. By the careful definition of the overall covariance structure, it is possible to handle cases when measurement errors are correlating or even have an interaction with the phylogenetic error. Then, the observer is left with a model-fitting problem, in which the task is to maximize the probability of data conditioned on the expected covariances. Therefore, parameter estimates from a best-fitting model with an error term that is composed of phylogenetic and measurement effects can be used to make inferences that are not confounded by these factors.

Such an approach has been taken forward by Ives et al. (2007), who derived statistical methods for the analysis of phylogenetically correlating data with within-species variation to investigate a broad array of evolutionary questions. Their entire methodology was built on the simple foundation that sampling variance can be added to the variance that is determined by the phylogenetic relationship of species (Martins and Hansen 1997). This scheme, on the one hand, can be applied to univariate models of evolution, when the interest is on ancestral states, rates of evolution and phylogenetic signal charactering the evolution of a single trait. Such a model can be depicted as (see also Eq. 7.10):

$$\mathbf{y} = \beta_a \mathbf{1} + \varepsilon_{\mathbf{S}} + \varepsilon_{\mathbf{M}} \tag{7.13}$$

$$\varepsilon_{\mathbf{S}} \sim \mathcal{N}(0, \sigma^2 \mathbf{C}), \tag{7.13a}$$

$$\varepsilon_{\mathbf{M}} \sim \mathcal{N}(0, \sigma_{within}^2 \mathbf{I}), \tag{7.13b}$$

where

$\mathbf{y}$ vector of the observed trait values (dimension: $N_{species} \times 1$),

β_a a scalar giving the expected value (i.e., ancestral state at the base of the tree) of the trait,

$\mathbf{1}$ vector of ones,

$\varepsilon_{\mathbf{S}}$ vector of variances caused by the phylogeny (dimension: $N_{species} \times 1$),

ε_M vector of measurement error variances (dimension: $N_{species} \times 1$),
C correlation structure defined by the phylogeny (dimension: $N_{species} \times N_{species}$),
I identity matrix (dimension: $N_{species} \times N_{species}$),
σ^2 the overall phylogenetically inherited variance (rate of evolution),
σ^2_{within} measurement or within-species variance vector (dimension: $N_{species} \times 1$).

This signifies that the total error term $\varepsilon = \varepsilon_S + \varepsilon_M$ depicts a multivariate normal distribution and has a covariance matrix $\sigma^2 C + \sigma^2_{within} I$. The phylogenetic component comes from a distribution with a zero mean and a variance that is described by the covariation matrix $\sigma^2 C$. This matrix is composed of σ^2 that represents the phylogenetically inherited variance (i.e., the rate of evolution), and C that stands for the correlation structure that can be defined by the length of shared branches on the phylogeny. On a similar vein, the measurement error component also follows a normal distribution with a zero mean and with a covariance matrix of $\sigma^2_{within} I$ that gives the diagonal matrix of measurement errors (or within-species variances). These divisions assume that the trait evolution follows the Brownian motion model and that measurement errors are uncorrelated. However, other evolutionary models can be accommodated by the appropriate translation of branch lengths into the C matrix, while correlated measurement errors can be treated by nonzero off diagonals in the $\sigma^2_{within} I$ (which then can be written as $\sigma^2_{within} M$). Elements of C and σ^2_{within} are given by the data (i.e., known phylogeny and standard errors around species-specific means), the only parameters that are unknown are the ancestral state (β_a) and the rate of evolution (σ^2). To estimate these parameters, Ives et al. (2007) describe a model-fitting iteration processes based on estimated generalized least squares (EGLS), maximum likelihood (ML) and restricted maximum likelihoods (REML). In addition, by determining the rate of evolution under the assumption of no phylogenetic structure in the data ($C = I$, assuming star phylogeny) and at the observed phylogeny, it also becomes possible to calculate the strength of the phylogenetic signal that is prevalent in the data in terms of Blombergs' K (Blomberg et al. 2003). Some procedures for characterizing univariate trait evolution by using the method by Ives et al. (2007) are given in the OPM.

Ives et al. (2007) also present a list of models for multivariate cases, when phylogenetic associations between traits are in the focus. Along this line, they make PGLS technique suitable for designs such as correlations, principal component analysis, multiple regression as well as reduced major axis regression for functional relationships (but Hansen and Bartoszek (2012) warn against this last application). The underling formula has a composition of

$$\mathbf{w} = \mathbf{b} + \varepsilon_S + \varepsilon_M \qquad (7.14)$$

where
$\mathbf{w}$ vector of species-specific tip values for traits $\mathbf{x}$ and $\mathbf{y}$ that are placed on top of each other (dimension: $2N_{species} \times 1$),

b   vector containing the ancestral states (β_a) for the two traits, with the first N_{species} elements being β_{a_x}, while the second N_{species} elements being β_{a_y},

$\boldsymbol{\varepsilon_S}$ phylogenetic error term vector (dimension: $2N_{\text{species}} \times 1$), in which the phylogenetic variance on **x** ($\boldsymbol{\varepsilon_{S_x}}$) is stacked on top of the phylogenetic variance on **y** ($\boldsymbol{\varepsilon_{S_y}}$),

$\boldsymbol{\varepsilon_M}$ measurement error vector (dimension: $2N_{\text{species}} \times 1$), in which the within-species variances of **x** ($\boldsymbol{\varepsilon_{M_x}}$) is stacked on top of the within-species variance vector for **y** ($\boldsymbol{\varepsilon_{M_y}}$).

The phylogenetic error term ($\boldsymbol{\varepsilon_S}$) has a joint covariance matrix, in which the diagonal blocks are $\sigma_x^2 \mathbf{C}$ and $\sigma_y^2 \mathbf{C}$, while the off-diagonal blocks are composed of $\sigma_x \sigma_y \mathbf{C}$ matrices that are multiplied by a linear combination of parameters that describes the association between traits **x** and **y** (i.e., r, correlation coefficient or β, regression slope). The error term $\boldsymbol{\varepsilon_M}$ is organized analogically, thus has a joint covariance matrix based on blocks of $\boldsymbol{\sigma_{\text{within}_x}^2}$ and $\boldsymbol{\sigma_{\text{within}_y}^2}$ measurement error variances in the diagonal, and matrices of $\boldsymbol{\sigma_{\text{within}_x} \sigma_{\text{within}_y}}$ representing the covariances in measurement errors scaled by a factor that is proportional to the strength of association between **x** and **y**. The unknown parameters in these models are β_{a_x}, β_{a_y}, σ_x^2, σ_y^2, and the coefficient r or β reflecting the strength of relationship between **x** and **y** originating from their correlated evolution. The procedures based on EGLS, ML, and REML approaches can be used for the estimation of these parameters and even can be flexibly extended to multivariate cases such as principal component analysis or multiple regression (i.e., when $\beta = (\beta_1, \beta_2 \ldots)$). Ives et al. (2007) recommend parametric bootstrapping (a procedure, in which estimated parameters are used to simulate a large number of datasets; then, the parameters are repeatedly re-estimated from each simulated data) to determine confidence interval around these estimates, which can be used for hypothesis testing (i.e., r or $\beta \neq 0$). (see examples in the OPM).

Emphasizing the importance of the discrimination between observation errors acting on the response and predictor variables in evolutionary regressions, Hansen and Bartoszek (2012) suggested an alternative PGLS model. In their appraisal, employing the above abbreviations, the general statistical equation can be written as:

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \cdots + \beta_m \mathbf{x}_m + \boldsymbol{\varepsilon_S} + \boldsymbol{\varepsilon_M} - \left(\beta_1 \boldsymbol{\varepsilon_{M_{x_1}}} + \cdots + \beta_m \boldsymbol{\varepsilon_{M_{x_m}}} \right) \quad (7.15)$$

where

y   vector of the dependent variable (dimension: $N_{\text{species}} \times 1$),

$\mathbf{x}_1 \ldots \mathbf{x}_m$ vectors for $1 \ldots m$ dependent variables (each with the dimension of $N_{\text{species}} \times 1$),

$\beta_0 \ldots \beta_m$ regression parameters including the intercept (β_0) and the slopes for each predictors ($\beta_1 \ldots \beta_m$),

$\boldsymbol{\varepsilon_S}$ residual phylogenetic error term vector,

ε_M measurement error vector for the dependent variable,
$\varepsilon_{M_{x_1}} \ldots \varepsilon_{M_{x_m}}$ measurement error vectors for each $(1\ldots m)$ predictor variable.

Therefore, the residual error in the model is composed of three components; the joint covariance matrix includes (1) a matrix that describes the model of evolution (in a form of $\sigma^2 \mathbf{C}$), (2) a matrix of raw observation variance in the response variable (in a form of $\sigma^2_{\mathbf{within_y}}$), and (3) a matrix representing the effects of measurement error in the predictor variables (a complicated variance structure relying on observation variances conditional on the observed values of the predictor variables) that is multiplied by the associated regression slopes (formally can be written as $\text{Var}[\beta\sigma^2_{\mathbf{within_x}}|\mathbf{X}]$). The effect of measurement errors in the predictors needs to be complex, because errors in one predictor will carry over and have an influence on coefficients that pertain to other predictors.

Hansen and Bartoszek (2012) argued that applying an EGLS procedure to obtain regression parameters from the above model results in more precise confidence intervals around the estimates, but it does not remove the downward biases that are caused by measurement errors around the predictors. In classical models of measurement errors (Madansky 1959; Fuller 1987; Buonaccorsi 2010), such attenuation bias can be corrected via a reliability ratio, K. This is defined as the ratio between the true and observed sum of squares of the predictor variable, with the former being estimated by subtracting the observational variance from the observed sum of square. Hansen and Bartoszek (2012) present equations for the calculation of reliability ratio when data are correlating due to the phylogenetic structure, thus the logic of using a correction factor on the regression slope via K can be adopted to aforementioned PGLS approach. They also include these calculations in their estimation procedure, in which they combine ML-based and EGLS approaches for the estimation of different parameters in their complex model. In the OPM, I show how these functions (as implemented in GLSME) work in practice.

The main difference between the method by Ives et al. (2007) and that of Hansen and Bartoszek (2012) is that the former approach does not strategically discriminate between the effect of measurement error around the predictor and response variables (these are lumped within the $\sigma^2_{\mathbf{within}}$ matrix). On the other hand, Hansen and Bartoszek (2012) uses separate matrices for the observation variance for the response $\sigma^2_{\mathbf{within_y}}$ and predictor $\sigma^2_{\mathbf{within_x}}$ variables, with the latter having complicated effects. Furthermore, strategies are different with regard how the two methods correct for attenuation bias (e.g., for this purpose Hansen and Bartoszek use the reliability ratio parameter).

Further Developments on the Phylogenetic Mixed Models

Implementing Lynch's original suggestion, Housworth at al. (2004) and Hadfield and Nakagawa (2010) brought the framework based on mixed modeling for the

study the evolution of traits into practice. Along this line, Hadfield and Nakagawa (2010) emphasized that animal models used to decompose variance components in quantitative genetics are built on a mathematical basis that is very similar to what is applied in phylogenetic mixed models. This analogy arises because the matrices that define the relationship between subjects, i.e., those that represent the phylogeny and the pedigree, can be brought into a structure that is formally equivalent. Based on this relationship, the quantitative genetics toolbox can be efficiently exploited for the decomposition of error structures in phylogenetic comparative studies (see more details in Chap. 11).

Accordingly, many test situations in interspecific comparative studies can be regarded as variations to the same underlining statistical foundation, the phylogenetic meta-analysis. In this context, the general mixed model for univariate questions can be written as (see also Eqs. 7.1, 7.8 and 7.9)

$$y_i = \beta_0 + a_i + e_i + m_i, \tag{7.16}$$

$$\mathbf{m} \sim \mathcal{N}(0, \sigma^2_{\mathbf{within}}\mathbf{I}) \tag{7.16a}$$

where

y_i	species-specific (or study-specific) effect size or trait value,
m_i	measurement error for species (or study) i (ith element of the $\mathbf{m}$ vector with dimension of $N_{\text{species}} \times 1$),
$\sigma^2_{\mathbf{within}}$	within-species variance caused by differences between individuals or species, the corresponding identity matrix ($\mathbf{I}$) that has a dimension of $N_{\text{species}} \times N_{\text{species}}$,
β_0, a_i, and e_i	intercept (ancestral state at the root of the phylogeny), the effect on species (or study) i that is caused by the common descent and the non-heritable residual component of variation (residual error), respectively, as in Eq. (7.8).

Chapter 11 shows how it can be extended into bi- or multivariate problems. The matrix of expected (co)variances among subjects caused by the random effects can be thus approached by a joint covariance matrix that is composed of the phylogenetic variance (again, in a form of $\sigma^2\mathbf{C}$), the measurement error variance (in a form of $\sigma^2_{\mathbf{within}}\mathbf{I}$ if such errors are known) and the residual variance ($\sigma^2_e\mathbf{I}$ assuming that residuals are homoscedastic). The effects that are lumped in the residual component may be important if deviations from the expected species-specific means are mediated by processes that act independently of phylogeny and within-species variances. Note that most of the above models can also be fitted to this general outline. For example, the method by Ives et al. (2007) considers a covariance structure, in which the phylogenetic and measurement error components are combined in the same additive fashion, and which is equivalent with the general phylogenetic meta-analysis model under the $\sigma^2_e = 0$ scenario. When species-specific traits could be completely described by the Brownian motion of

evolution, Felseinstein's (2008) model can also be brought into a similar structure with the main difference that σ^2_{within} is common to all species and remains an unknown quantity.

To determine the unknown parameters in the model (e.g., β_0 and σ^2), previous models based on PGLS or independent contrasts use estimation procedures available in ML, EGLS, REML, or EM algorithms. However, for these approaches, it becomes challenging to deal with non-Gaussian distributions as well as with missing information regarding the species-specific trait values, within-species variances, and phylogenetic resolutions. To overcome these shortcomings, Hadfield and Nakagawa (2010) proposed a method based on Markov chain Monte Carlo (MCMC) simulation that can be accommodated to a wide range of phylogenetic questions and data types. MCMC can be used to fit the general model:

$$l = \mathbf{W}\boldsymbol{\theta} + \mathbf{e}, \tag{7.17}$$

where

l a latent variable that provides the link function (e.g., Poisson and exponential) to the values of the $\mathbf{y}$ response variable,

$\mathbf{w}$ design matrix of predictor variables $\mathbf{x}_1 \ldots \mathbf{x}_m$,

$\boldsymbol{\theta}$ vector of fixed and random effects, and

$\mathbf{e}$ vector of residuals.

Estimation protocols are needed to obtain l, $\boldsymbol{\theta}$, the variance components σ^2 (corresponding to the distribution of phylogenetic random effect included in $\boldsymbol{\theta}$), and σ^2_e (corresponding to the distribution of residual error $\mathbf{e}$). The estimation of l is achieved through a Metropolis–Hasting iteration process (Metropolis et al. 1953; Hastings 1970), while θ and the variance components are approximated by Gibbs sampling (Geman and Geman 1984). By including a matrix of within-species variances into the definition of $\boldsymbol{\theta}$, such sources of variation can be flexibly incorporated into models that represent various evolutionary questions. The obtained parameter estimates from such models thus can be regarded as being independent of the confounding effect of measurement error. Check the OPM for an example based on the *R* package *MCMCglmm*.

Methods Based on the Evaluation of Likelihood Surfaces

Beside the problems posed by non-Gaussian distribution and missing data, another limitation can appear in classical approaches to phylogenetic comparative questions. Specifically, they are constrained to use particular assumptions about the mode of trait evolution. The evolutionary models that are traditionally considered are the Brownian motion and the Ornstein–Uhlenbeck process, whose likelihood functions are mathematically tractable for parameter estimations. For example, due to analytical convenience, the above models generally accommodate the Brownian fashion of evolution during the translation of branch length into the

covariance matrix (i.e., elements of **C** are proportional to branch lengths). Some departures from this standard can be achieved by the appropriate adjustment of the formula, but more complicated models of evolution, such as branch-specific directional selection requiring the estimation of large number of parameters, cannot be tracked in this way.

To surmount this obstacle, Kutsukake and Innan (2012, see also Chap. 17) introduced a method that is able to deal with more complex and realistic modes of evolution. Instead of forming mathematical formulas to describe the relationship between known data and the parameters of interest, they advocate the simulation of data at different parameter settings to examine how well such simulation results coincide with the observed data. By simulating a large number of data at various parameter sets representing different hypotheses about patterns of evolutions, we only need to examine the likelihood of each set based on the joint probability of the observed and simulated data. Parameter combinations that provide the highest likelihood can be used for making evolutionary inferences. If this ML estimation is supported by approximate Bayesian computation (ABC, see Chap. 17), the algorithm can take into account prior information on the expected distributions for all parameters that can result in increased power.

An important flexibility of the above simulation-based method is that it can also incorporate intraspecific variation. Although the simulation process focuses on phenotypic trait values as estimated at the tips of the phylogeny, and the simulated and the observed data are compared at the level of species, the likelihood function can be adjusted for patterns of phenotypic variation accumulating within species. Accordingly, the joint probability of observed and simulated data can be extended to include species-specific trait values as well as standard deviations around them. In fact, not only normal distributions, but also other forms of within-species distributions can be considered. Consequently, if data are available on how the data are spread within species, such information can be efficiently incorporated in the estimation of likelihood surfaces of model parameters. Focusing on the comparison of three evolutionary models, Kutsukake and Innan (2012) present an example analysis for the estimation of ancestral statesbased on phenotypic data (mean and standard deviation) that are sorted at the level of species. However, their logic can be tailored to various evolutionary questions, and equivalent approaches can be designed for multitrait evolution, when the interest is to obtain parameters to describe patterns of correlated evolution.

Another likelihood-based method to incorporate intraspecific variation has also been developed in the Bayesian framework that relies on MCMC methods (Revell and Reynolds 2012). The main difference between this and most of the above-discussed comparative methods is that Bayesian method uses trait data that are broken down to individuals, while species-specific trait values are not needed as an input (note that merely the independent contrast approach relies on similar requirements). Therefore, only the phylogeny, individual values and evolutionary models are treated as known (Revell and Reynolds 2012 considered Brownian process for their description, but other evolutionary models can also be envisaged). Parameters that express species means and variances are estimated together with

the parameters of the evolutionary model from their joint probability distribution. Therefore, such an approach can account for the possibility that processes involved in the considered evolutionary model and its parameters can affect species means and variances. The simulations accompanying the methodology of Revell and Reynolds (2012) indicated that true species means are not necessarily the same as the arithmetic mean of individual-specific values, especially when intraspecific variance is relatively high. This suggests that accounting for the tree and the modes of trait evolution may be warranted when inferring about tip values at the species level from within-species samples. This does not only apply to the means but also to variances, as patterns of the intraspecific distribution of traits may differ between species in a phylogenetically determined fashion. In the OPM, I demonstrate the use of this Bayesian method focusing on functions available in *phytools*.

The philosophy of the two above methods (Kutsukake and Innan 2012; Revell and Reynolds 2012) is very similar in the sense that they both revolve around a maximization of a likelihood function for the proposed set of evolutionary parameters. To obtain this, one needs to derive the probability of data conditioned on the means and variances of species as well as on the parameters of the underlying evolutionary model and the phylogeny. While Kutsukake and Innan (2012) evaluate the surface of likelihoods through processes of data simulation at different combinations of parameters, Revell and Reynolds (2012) put forward an MCMC-based Bayesians algorithm to obtain the joint posterior probability distributions of parameters. In the latter procedure, propositions for parameter values are being made at each node of the chain, and these propositions are accepted proportional to their likelihood.

There are also differences in how intraspecific variance is incorporated in the likelihood function. The simulation-based method assumes that this is a known property and thus the probability function of data given the simulated values can be simply adjusted. On the other hand, in the Bayesian approach, within-species variations are unknown, thus two different probabilities are needed for the formation of likelihood. The first part gives the probability that the proposed mean data arose from the model of evolution and phylogenetic tree, while the second part describes the probability that observed individual-specific data conditioned on the proposed species-specific means and variances. The advantages of the different strategies applied in the above two methods might be exploited depending on the question and data at hand. For example, if the biological hypothesis under testing is related to the evolution of species-specific trait and within-species variance is only regarded as a confounder, the model by Kutsukake and Innan (2012) may be appropriate. However, if we have individual-specific values, we can investigate interesting questions about the evolution of within-species variances.

Adjusting for Unequal Within-Species Sample Sizes

To correct for heterogeneity in sample sizes among subjects, one can use weighted regressions (Draper and Smith. 1981; Neter et al. 1996), in which each data point

is given an emphasis according to the corresponding sample size. Such a weighting approach can also be adopted for the phylogenetic framework, but it has seen little test in that context. Given that standard error reciprocally scales with sample size (Box 7.1), the above methods (e.g., PGLS, phylogenetic mixed models) using information on error variances can be supplied with $1/n_i$ as an estimate of within-species variance component. Such an analysis will give results that are adjusted for differences in sample size. If both standard errors and sample sizes are provided, the within-species variance that is corrected for sample size can be calculated based on the regression of standard error against the sample size, which can be subsequently used in a measurement error model. Some examples are shown in the OPM.

Another way to deal with non-constant sample size is to apply data imputation method that produces estimates for cases when information is unavailable. Variation in sample size can be considered as a consequence of missing information for some individuals (Garamszegi and Møller 2011). Various approaches are available to input missing data (Nakagawa and Freckleton 2008) in order either to equalize within-species sample size by augmenting individual-specific data or even to simulate data for species for which no data are available at all. Unfortunately, such imputation methods have been rarely exploited in the phylogenetic contexts (see Fisher et al. 2003).

7.4.4 Interpreting Phylogenetic Results in Light of Within-Species Variation

Results from the above exercises should be interpreted carefully. It has been noted, for example, that different approaches may provide somewhat different outcomes (e.g., Ives et al. 2007). Hence, repeating the analyses by using alternative methods if these are available may help establish our confidence in the validity of detected patterns. If discrepancies are found, we may need to revisit the assumptions of different models to check whether these were violated. Furthermore, the visual inspection of data can also enhance the interpretations. For instance, the types of graphics presented in this chapter show error ranges or sample sizes around the species-specific estimates and also illustrate the results with and without accounting for measurement errors (Figs. 7.1 and 7.2).

How does controlling for within-species variance change the results compared to the situation when species-specific mean values are assumed to have no errors? Does this difference correspond well with the estimated trait repeatabilities (i.e., high repeatability should cause only minor difference between different outcomes)? Answers to such questions may help elucidate the validity of the findings. The inspection of the estimated parameters can also be informative. In general, as discussed above, we should expect that a control for within-species variance increases phylogenetic signal in the data, while the regression slope or correlation

coefficient strengthens after removing the attenuation bias that the measurement error causes. Finally, we can also check the model fit statistics to verify whether the model accounting for within-species variance offers better fit to the data.

7.5 Discussion

In this chapter, I investigated issues about the importance of within-species variance in phylogenetic comparative studies that generally focus on species-specific means as a unit of analysis. Evidence from simulation studies suggests that this focus on interspecific patterns should not inherently imply that intraspecific patterns are to be ignored. The potential problems posed by measurement errors around species-specific means do not only have consequences for how we analyze data, but also for how we design and collect data for comparative studies, how we interpret phylogenetic findings, and ultimately, how we think about evolution.

Warnings about the importance of the incorporation of within-species variance into analyses at the across-species level have emerged in the recent literature, which may likely demarcate avenues for the development of the methodology in the future. However, I would argue that before such statistical developments take place, empirical studies are needed that confirm the biological relevance of the appropriate control methods. Lessons about the use of phylogenetic control teach us that although interspecific data are unavoidably structured by common ancestry, the phylogenetic control is warranted only if the data at hand require so (Freckleton 2009; Revell 2010). Similarly, the smart application of measurement error models that also involves biological considerations and a closer look into the data should be preferred over the blind submission of subsequent comparative datasets to complex analyses with intraspecific variation (note that more complex models usually require fulfilling more assumptions). Accordingly, in spite of the fact that in theory it seems necessary to deal with the confounding effects arising from within-species distributions, in practice it may appear that the available data represent a case when variation below the species level is negligible, and when classical comparative methods perform with high confidence as well. I suspect that this situation will call for the performance of diagnostics statistics in most of the studies rather than full exploitation of phylogenetic approaches that account for measurement error.

Thinking based on biological motivation may also direct us into a fascinating research direction. So far, statistical considerations implied that within-species variation is a somewhat unwanted side effect that we should get rid of in the comparative analysis. This echoes the philosophy that was applied in the early days of phylogenetic methodology, i.e., when the phylogenetic structure in the data was regarded as something that should be removed from the data, for example via the use of independent contrasts. Only later progress realized that the phylogenetic trees in fact may be incorporated into the analyses in a more beneficial fashion that allows making inferences about the modes of evolution (see Chap. 1). In a similar

vein, from the current state of the art, subsequent research may also recognize that measurement error in a statistical sense may hold interesting information from the biological perspective (see Sect. 7.2) that can be exploited fruitfully. Therefore, methods that do not only control for and factor out intraspecific variation, but can also deal with its evolutionary relevance may open new dimensions. For instance, within-species variance itself can be subject to selection, thus incorporating such information into the phylogenetic comparative study as a covariate rather than handling it as a confounder may provide interesting results. As an example, in a study of flight initiation distance in birds, Møller and Garamszegi (2012) found that within-species variance of this trait can be shaped by ecological factors, suggesting that these are not just random variation around a species-specific mean. Another toolbox that holds promises toward the same direction is the phylogenetic meta-analysis (see Chap. 11), which brings effect sizes together with their confidence intervals into the focus. In such an approach, each species-specific effect size represents the strength (and direction) of a particular relationship between two traits, while confidence intervals describe the precision of the mean effect size estimate based on the underlying within-species sample size. The meta-analytic study of phylogenetically structured effect sizes is basically a comparative problem that also considers within-species variance around species-specific estimates. The exploitation of such methods for investigating evolutionary questions awaits further progress. These may be interesting, for example, when traits can show correlations at both the within- and between-species level (see Fig. 7.2), and the aim is to explain why some species display strong relationships, while others weak relationship between two phenotypic traits.

The approaches discussed in this chapter generally assume that individual- or population-specific measurements are independent of each other and thus depict no further phylogenetic or other hierarchical structure within species (i.e., they can be visualized on the phylogeny as forming a star polytomy with zero branch length). However, such an assumption may not be necessarily true, especially when within-species variance arises from variations between populations. In fact, populations of the same species can have a certain evolutionary history, thus they cannot be regarded as phylogenetically independent replicates, such as data from different individuals (see Edwards and Kot 1995 who first used phylogenetic comparative methods on intraspecific data). Moreover, populations are structured in space that has consequences for migration and gene flow so that phenotypes in one locality are affected by processes acting in other neighboring localities. Therefore, inferences made from across-population patterns need to consider statistical issues about non-independence at least due to two factors: phylogeny and gene flow. Felsenstein's group described various methods that are able to quantify evolutionary patterns across multiple populations within a single species (Stone et al. 2011).

Such methods may, however, be developed further, and combined with other comparative methods to partition variances and evolutionary patterns acting at the between-population and the between-species levels (e.g., by relying on the mixed model framework). For example, it might be straightforward to first estimate

species-specific trait values and variances over multiple populations by incorporating the effects of phylogeny and gene flow sensu Stone et al. (2011) and then subsequently use such species-specific estimates in an interspecific comparison. Moreover, it might be straightforward to make distinction between cases when populations should be treated as separate entries in the analysis and when they should be pooled as one species. This might, of course, depend on several factors such as data availably, the biological question, the relative importance of gene flow, and phylogenetic constraints.

I envisage there being great potential for further development of comparative methods incorporating measurement error along various other lines. Most of the advancements have been made so far correspond to situations when the correlated evolution of traits is of interest. However, there might be other phylogenetic problems and designs that also warrant considerations about within-species variance. In practice, we might also need specific methods that are able to deal with non-normal or skewed within-species distributions, count data as well as with missing data. Furthermore, so far there is not a strong distinction between different sources of within-species variance in the statistical approaches. Therefore, it may prove useful to derive methods that can separate or combine variance components that originate from instrumental errors, within- or between-individual variations, and fluctuations across populations. Finally, there might be a scope for amalgamating methods that implement uncertainty in estimating tip values and that consider measurement errors in a phylogenetic tree (de Villemereuil et al. 2012).

References

Adolph SC, Hardin JS (2007) Estimating phenotypic correlations: correcting for bias due to intraindividual variability. Funct Ecol 21(1):178–184

Arnold C, Nunn CL (2010) Phylogenetic targeting of research effort in evolutionary biology. Am Nat 176:601–612

Ashton KG (2004) Comparing phylogenetic signal in intraspecific and interspecific body size datasets. J Evol Biol 17:1157–1161

Blomberg S, Garland TJ, Ives AR (2003) Testing for phylogenetic signal in comparative data: behavioral traits are more laible. Evolution 57:717–745

Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, White JSS (2009) Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol Evol 24:127–135

Bollen KA (1989) Structural equations with latent variables. Wiley, New York

Buonaccorsi JP (2010) Measurement error: models, methods, and applications. Chapman and Hall, New York

Caro TM, Roper R, Young M, Dank GR (1979) Inter-observer reliability. Behaviour 69:303–315. doi:10.1163/156853979x00520

Chesher A (1991) The effect of measurement error. Biometrika 78(3):451–462. doi:10.1093/biomet/78.3.451

Cheverud JM, Dow MM, Leutenegger W (1985) The quantitative assessment of phylogenetic constraints in comparative analyses: sexual dimorphism of body weight among primates. Evolution 39:1335–1351

Christman MC, Jernigan RW, Culver D (1997) A comparison of two models for estimating phylogenetic effect on trait variation. Evolution 51(1):262–266. doi:10.2307/2410979

Cornillon PA, Pontier D, Rochet MJ (2000) Autoregressive models for estimating phylogenetic and environmental effects: accounting for within-species variations. J Theor Biol 202(4):247–256. doi:10.1006/jtbi.1999.1040

de Villemereuil P, Wells JA, Edwards RD, Blomberg SP (2012) Bayesian models for comparative analysis integrating phylogenetic uncertainty. BMC Evol Biol 12: doi:10.1186/1471-2148-12-102

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J Roy Stat Soc Ser B Methodol 39(1):1–38

Doughty P (1996) Statistical analysis of natural experiments in evolutionary biology: comments on recent criticisms on the use of comparative methods to study adaptation. Am Nat 148:943–956

Draper NR, Smith H (1981) Applied regression analysis. Wiley, New York (2nd edn)

Edwards SV, Kot M (1995) Comparative methods at the species level: geographic variation in morphology and group size in Grey-crowned Babblers (*Pomatostomus temporalis*). Evolution 49:1134–1146

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

Felsenstein J (2002) Contrasts for a within-species comparative method. In: Slatkin M, Veuille M (eds) Modern developments in theoretical population genetics. Oxford University Press, Oxford, pp 118–129

Felsenstein J (2004) Inferring phylogenies. Sinauer Associates, Sunderland

Felsenstein J (2008) Comparative methods with sampling error and within-species variation: contrasts revisited and revised. Am Nat 171(6):713–725

Fisher DO, Blomberg SP, Owens IPF (2003) Extrinsic versus intrinsic factors in the decline and extinction of Australian marsupials. Proc R Soc Lond Ser B-Biol Sci 270(1526):1801–1808

Freckleton RP (2009) The seven deadly sins of comparative analysis. J Evol Biol 22(7):1367–1375

Fuller WA (1987) Measurement error models. Wiley, New York

Garamszegi LZ, Møller AP (2010) Effects of sample size and intraspecific variation in phylogenetic comparative studies: a meta-analytic review. Biol Rev 85:797–805

Garamszegi LZ, Møller AP (2011) Nonrandom variation in within-species sample size and missing data in phylogenetic comparative studies. Syst Biol 60:876–880

Garamszegi LZ, Møller AP (2012) Untested assumptions about within-species sample size and missing data in interspecific studies. Behav Ecol Sociobiol 66:1363–1373

Gelman A, Hill J (2007) Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, Cambridge

Geman S, Geman D (1984) Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. Ieee Transactions on Pattern Analysis and Machine Intelligence 6(6):721–741

Gittleman JL, Kot M (1990) Adaptation: statistics and a null model for estimating phylogenetic effects. Syst Zool 39(3):227–241. doi:10.2307/2992183

Grafen A (1989) The phylogenetic regression. Philos Trans R Soc B 326:119–157

Hadfield JD, Nakagawa S (2010) General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. J Evol Biol 23:494–508

Hansen TF, Bartoszek K (2012) Interpreting the evolutionary regression: the interplay between observational and biological errors in phylogenetic comparative studies. Syst Biol 61:413–425

Harmon LJ, Losos JB (2005) The effect of intraspecific sample size on type I and type II error rates in comparative studies. Evolution 59:2705–2710

Hastings WK (1970) Monte carlo sampling methods using markov chains and their applications. Biometrika 57(1):97–109. doi:10.2307/2334940

Housworth EA, Martins EP, Lynch M (2004) The phylogenetic mixed model. Am Nat 163:84–96

Ives AR, Midford PE, Garland T (2007) Within-species variation and measurement error in phylogenetic comparative methods. Syst Biol 56(2):252–270

Judge GG, Griffiths WE, Hill RC, Lutkepohl H, Lee T-C (1985) The theory and practice of econometrics. Wiley, New York

Kreft IGG, de Leeuw J, Aiken LS (1995) The effect of different forms of centering in hierarchical linear models. Multivar Behav Res 30:1–21

Kutsukake N, Innan H (2012) Simulation-based likelihood approach for evolutionary models of phenotypic traits on phylogeny. Evolution (in press)

Lessells CM, Boag PT (1987) Unrepeatable repeatabilities: a common mistake. Auk 104:116–121

Lynch M (1991) Methods for the analysis of comparative data in evolutionary biology. Evolution 45(5):1065–1080

Madansky A (1959) The fitting of straight lines when both variables are subject to error. J Am Stat Assoc 54:173–205

Manisha S (2001) An estimation of population mean in the presence of measurement errors. J Indian Soc Agric Stat 54:13–18

Martins EP (1994) Estimating the rate of phenotypic evolution from comparative data. Am Nat 144:193–209

Martins EP (2004) COMPARE, version 4.6b. Computer programs for the statistical analysis of comparative data. Distributed by the author at http://compare.bio.indiana.edu/, Department of Biology, Indiana University, Bloomington IN

Martins EP, Hansen TF (1997) Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. Am Nat 149:646–667

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. J Chem Phys 21:1087–1092

Møller AP, Garamszegi LZ (2012) Between individual variation in risk taking behavior and its life history consequences. Behav Ecol 23:843–853

Monkkonen M, Martin TE (2000) Sensitivity of comparative analyses to population variation in trait values: clutch size and cavity excavation tendencies. J Avian Biol 31(4):576–579. doi:10.1034/j.1600-048X.2000.310417.x

Nakagawa S, Freckleton R (2008) Missing inaction: the dangers of ignoring missing data. Trends Ecol Evol 23:592–596. doi:10.1016/j.tree.2008.06.014

Nakagawa S, Schielzeth H (2010) Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. Biol Rev 85:935–956

Neter J, Kutner MH, Nachtsheim CJ, Wasserman W (1996) Applied linear statistical models. Irwin, Chicago

Paradis E (2011) Analysis of phylogenetics and evolution with R, 2nd edn. Springer, Berlin

Purvis A, Rambaut A (1995) Comparative analysis by independent contrasts (CAIC): an Apple Macintosh application for analysing comparative data. Comput Appl Biosci 11:247–251

Purvis A, Webster AJ (1999) Phylogenetically independent comparisons and primate phylogeny. In: Lee PC (ed) Comparative primate socioecology. Cambridge University Press, Cambridge, pp 44–70

Reed GF, Lynn F, Meade BD (2002) Use of coefficient of variation in assessing variability of quantitative assays. Clin Diagn Lab Immunol 9(6):1235–1239. doi:10.1128/cdli.9.6.1235-1239.2002

Revell LJ (2010) Phylogenetic signal and linear regression on species data. Methods Ecol Evol 1(4):319–329. doi:10.1111/j.2041-210X.2010.00044.x

Revell LJ, Reynolds RG (2012) A new Bayesian method for fitting evolutionary models to comparative data with intraspecific variation. Evolution 66(9):2697–2707. doi:10.1111/j.1558-5646.2012.01645.x

Ricklefs RE, Starck JM (1996) Applications of phylogenetically independent contrasts: a mixed progress report. Oikos 77(1):167–172

Snijders TAB, Bosker RJ (1999) Multilevel analysis—an introduction to basic and advanced multilevel modelling. Sage, London

Sokal RR, Rohlf FJ (1995) Biometry, 3rd edn. W. H. Freeman and Co, New York

Stone GN, Nee S, Felsenstein J (2011) Controlling for non-independence in comparative analysis of patterns across populations within species. Philos Trans R Soc Lond B Biol Sci 366:1410–1424

Taper ML, Marquet PA (1996) How do species really divide resources? Am Nat 147:1072–1086

van de Pol MV, Wright J (2009) A simple method for distinguishing within- versus between-subject effects using mixed models. Anim Behav 77(3):753–758

Chapter 8
An Introduction to Phylogenetic Path Analysis

Alejandro Gonzalez-Voyer and Achaz von Hardenberg

The questions addressed by macroevolutionary biologists are often impervious to experimental approaches, and alternative methods have to be adopted. The phylogenetic comparative approach is a very powerful one since it combines a large number of species and thus spans long periods of evolutionary change. However, there are limits to the inferences that can be drawn from the results, in part due to the limitations of the most commonly employed analytical methods. In this chapter, we show how confirmatory path analysis can be undertaken explicitly controlling for non-independence due to shared ancestry. The phylogenetic path analysis method we present allows researchers to move beyond the estimation of direct effects and analyze the relative importance of alternative causal models including direct and indirect paths of influence among variables. We begin the chapter with a general introduction to path analysis and then present a step-by-step guide to phylogenetic path analysis using the *d*-separation method. We also show how the known statistical problems associated with non-independence of data points due to shared ancestry become compounded in path analysis. We finish with a discussion about the potential effects of collinearity and measurement error, and a look toward possible future developments.

Both authors contributed equally to this work.

A. Gonzalez-Voyer (✉)
Conservation and Evolutionary Genetics Group, Estación Biológica de Doñana (EBD-CSIC), Av. Américo Vespucio SN, 41092 Sevilla, Spain
e-mail: alejandro.gonzalez@ebd.csic.es

A. von Hardenberg
Alpine Wildlife Research Centre, Gran Paradiso National Park, Degioz 11, 11010 Valsavarenche, Aosta, Italy
e-mail: achaz.hardenberg@pngp.it

8.1 Phylogenetic Linear Models: Drawbacks and Limitations When Analyzing the Influence of Multiple Variables

Because comparative biologists address questions related to long-term processes, they are faced with an important practical obstacle: the time necessary to produce evolutionary change. Hence, as with many problems in ecology, evolution, and behavior, the questions addressed by comparative biologists are often impervious to experimental approaches and alternative methods have to be adopted. Phylogenetic comparative methods employ the results from replicated "natural experiments" across multiple extant taxa as the data with which to test evolutionary hypotheses. Repeated associations between putative functional traits and environmental variables (proxies for a selective regime) or among traits are taken as supporting the evolutionary hypothesis. However, although the approach is potentially a very powerful one, given that comparisons generally involve numerous species and hence span long periods of evolutionary change (Freckleton 2009), comparative biologists are constrained in the inferences or conclusions they can draw from their results. Correlations between traits or between traits and the environment in extant taxa do not address the question of evolutionary origin (Martins 2000). Indeed, an important limitation when dealing with processes having occurred in the distant past is that there is no information about the conditions during most of the evolutionary history of the process being analyzed. Therefore, although there is a relationship between traits in extant taxa and current environments, this does not necessarily mean that there was a relationship between traits and the environmental conditions when the adaptation arose (Martins 2000). Furthermore, correlations between traits and the environment or between traits do not necessarily imply that the environment or trait is the driving force for the observed phenotypic changes (Martins 2000). Indeed, all correlative data have the inherent limitation that there is no way to determine causality. Nonetheless, a comparative method does exist allowing researchers to determine the order of evolutionary transitions (contingency) in correlated discrete traits (Pagel and Meade 2006).

These caveats not withstanding, there are also limitations regarding inferences researchers can make about their results due to limitations of the most commonly employed statistical methods. Currently, when testing hypotheses about associations between traits or traits and the environment, the method most often employed by comparative biologists is based on linear models. Phylogenetic independent contrasts or phylogenetic generalized least squares (PGLS) methods allow to analyze covariation between traits or traits and the environment, controlling for non-independence of data points (correlated residuals) due to shared ancestry (Felsenstein 1985; Grafen 1989; Martins and Hansen 1997). In addition, PGLS allows to combine continuous and discrete traits in a single model without the need to code dummy variables as well as allowing for different models of trait evolution to be incorporated in the analyses (Martins and Hansen 1997; see Chaps. 5 and 6).

However, both methods present similar limitations, which are the same as those of traditional linear models. First, only a single-dependent variable can be analyzed at a time, although a more realistic reflection of the complexity of the multivariate relationships currently analyzed by comparative biologists would allow for simultaneous exploration of the effects of a number of predictor variables on a number of different outcomes. Second, in multivariate linear models, a particular variable can either be a predictor or a response; however, a particular phenotypic trait can be responding, for example, to the influence of the environment and in turn be itself the cause of changes in a second phenotypic trait, hence a single trait can be both a response and a predictor. In order to overcome these limitations of traditional multivariate linear models, path analysis was developed. Confirmatory path analysis (and structural equation modeling) is an extension of multiple regression, but it is superior to ordinary regression analysis in that it allows researchers to move beyond the estimation of direct effects and analyze the relative importance of alternative causal models including direct and indirect paths of influence among variables. In von Hardenberg and Gonzalez-Voyer (2013), we introduced phylogenetic path analysis (PPA), integrating PGLS with the *d-separation* method for path analysis developed by Shipley (2000a). The proposed method allows researchers to harness the power of path analysis to disentangle cause–effect relationships among variables with data leading to correlated residuals due to shared ancestry. In this chapter and in the online practical material (hereafter OPM) available at www.mpcm-evolution.org, we provide further information and a detailed tutorial about how to perform PPA using the open source statistical language R (R Development Core Team 2013).

8.2 The Philosophy of Path Analysis

Correlation does not imply causation. Back in our undergraduate statistics classes, we were all taught this scientific mantra (Fig. 8.1). This statement is so deeply embedded in our modern scientific culture that it even deserved its own Wikipedia page.[1] Indeed, it is undeniable that if *A* is related to *B*, this does not imply that *B* is caused by *A*, or that *A* is caused by *B*. Both variables may, for example, be caused by a third confounding variable *C*. Some simple examples will elucidate this point: A highly significant correlation exists between the number of breeding pairs of storks (*Ciconia ciconia*) and human birthrates in Europe (Matthews 2000). Does this imply that storks deliver babies? Another study suggests that scientific productivity (measured as the number of citations) of ecologists is inversely correlated with per capita beer consumption (Grim 2008). Does this mean that beer drinking is detrimental to your scientific career? If you are not willing to give up your passion for beer, you may nonetheless be able to compensate eating lots of

---

[1] http://en.wikipedia.org/wiki/Correlation_does_not_imply_causation Retrieved June 4, 2014
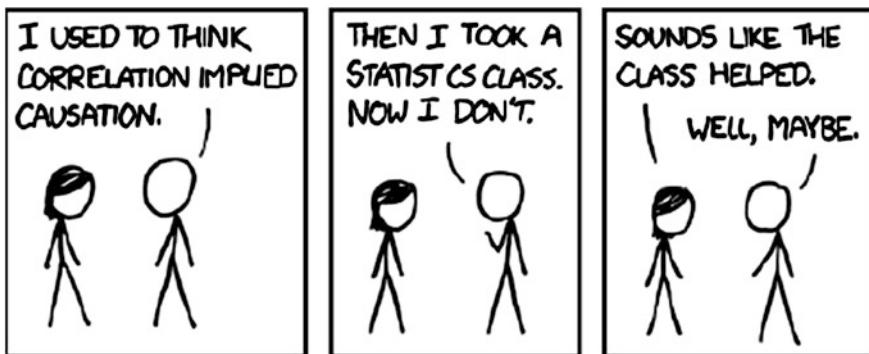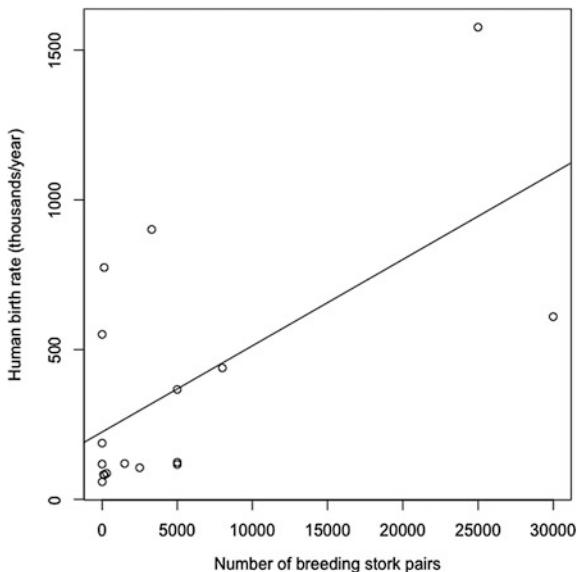
Fig. 8.1 Courtesy of XKCD (Distributed under a creative commons attribution-noncommercial 2.5 license)

chocolate. This at least is what a recent study in the New England Journal of Medicine suggests (Messerli 2012). The study shows a significant correlation between per capita chocolate consumption and the number of Nobel laureates per 10 million population in each country. All of these causal claims can easily be dismissed, taking into account other possible common causal variables, if not simply by logic. The main problem with the above-mentioned studies is that they are based on observational data rather than on controlled or randomized experiments (Fisher 1926), which are the commonly accepted scientific methods to infer causality. It would be great to be always able to use proper randomized or controlled experiments in all our studies, but this is obviously not possible, particularly in the case of comparative studies, where the unit of analysis is estimates of trait values for diverse species.

Correlation does not imply causation. This is what we have so dutifully learned. But is this completely true? Actually no. Indeed, without being afraid of saying a heresy, we can claim that *correlation always implies an underlying, unresolved causal structure* (Shipley 2000b). If we can rule out that the correlation between two variables is simply due to chance, there must be something that causes this relationship directly or indirectly through some other variables, even if we cannot necessarily identify the causes. The causal structure behind this correlation is indeed said to be *unresolved* because we cannot know, from the single correlation we can observe, how this correlation structure is built. Let us take a closer look at the "baby-delivering storks" data of Matthews (2000). The original data are available in the OPM available at http://www.mpcm-evolution.com as a "comma-separated values" (CSV) file. A tutorial showing how this data was analyzed and plotted using the open source statistical language R (R Development Core Team 2013), is also available in the aforementioned Web site. A quick glance at Fig. 8.2 strongly suggests that there is a relationship between number of storks and human births. Indeed, there is a significant relationship between the number of stork pairs and human birthrate with a *p* value of 0.008.

Fig. 8.2 Relationship between the number of breeding pairs of storks (*Ciconia ciconia*) and human birthrates in European countries (data from Matthews 2000)

Of course, none of us seriously believes that this means that storks really deliver babies[2]! Everybody will most likely agree that there is some other confounding variable which is the real, direct, or indirect cause of both the number of breeding stork pairs and human birthrates. As we said: *correlation always implies an underlying, unresolved causal structure*. Theoretically, we could test whether storks actually deliver babies doing a randomized experiment, for example, keeping constant the number of breeding stork pairs over a random selection of countries[3] in order to physically control for the variability in the number of stork pairs and at the same time excluding the effect of other possible confounding variables thanks to the randomization. However, this would undeniably be a very large scale and impractical experiment, not considering the moral implications it would have! What we can however do, if we have other factors which we suspect to be the true cause behind both human birthrates and the number of stork pairs, is to statistically control for the variability in these factors and thus see whether, controlling for the supposed common cause (in statistical jargon we would say: conditioning on it) the relationship between the number of storks and human birthrates still holds. We can try this using one of the other variables available in the dataset: area, which represents the surface size in squared kilometers of each country. It is indeed reasonable to think that larger countries host a higher number of stork pairs and at the same time have higher human birthrates, possibly indirectly through some other unmeasured variable. Indeed, there appears to be a very

---

[2] If you do, you can stop reading here!

[3] By hunting or, less drastically, translocating excess pairs from one country to an other.
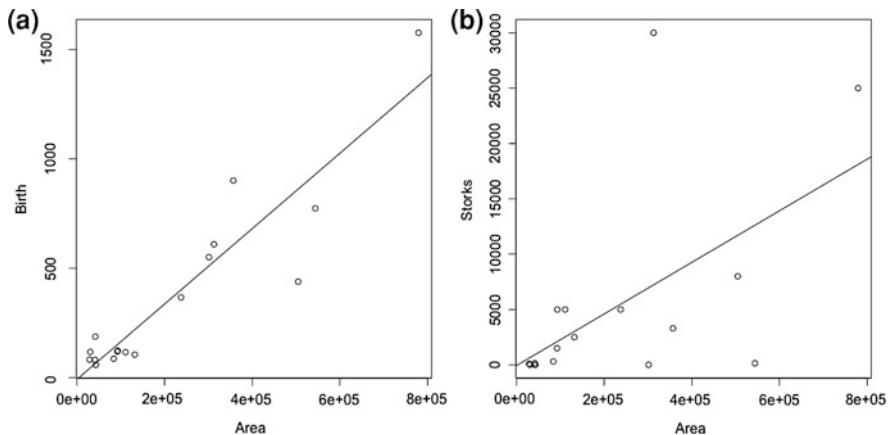
Fig. 8.3 **a** Relationship between human birthrate and country area; **b** relationship between the number of breeding pairs of storks and country area (data from Matthews 2000)

strong relationship between human birthrate and country area (Fig. 8.3a)! We can do the same for the relationship between the number of stork pairs and area. In this case also the relationship (Fig. 8.3b), even thought not as strong, is significant ($p = 0.0148$). We can now go back to our first linear model of the relationship between number of stork pairs and birthrates and statistically control for the effect of the confounding variable area, simply including this variable in the model transforming it into a multiple linear regression model of the kind[4]: Birth $\sim$ Area + Storks. Where Birth = birthrate, Area = country area, and Storks = number of stork pairs. With this model, while the effect of area on birthrate is highly significant ($p = 6.62e{-}06$), including this variable drastically changed the significance of the effect of the number of stork pairs to an unimpressive p value of 0.307 compared to the p value of 0.008 we obtained previously without conditioning on area! Technically, what we did is test the partial regression coefficient of the effect of the number of stork pairs on birthrate statistically controlling for the confounding effect of area and thus testing the statistical independence of the number of stork pairs from human birthrate. Is this enough to be able to claim that area is thus the common cause of both the number of breeding stork pairs and human birthrate? Sadly no. Indeed statistically, even if not logically, the result of this partial regression model may imply at least one alternative causal structure besides the above-mentioned hypothesis: The number of stork

---

[4] Note that here and in the rest of this chapter, we use the modified Wilkinson-Rogers notation for linear models (Wilkinson and Rogers 1973) widely used in statistical languages such as R. In this notation, the intercept is implicit and the tilde ($\sim$) separates the left-hand side from the right-hand side of the equation.

pairs may indirectly influence human birthrate through the area of countries.[5] While this alternative hypothesis does not necessarily make any logical sense in this case,[6] statistically, in the absence of further information, we cannot distinguish it from the hypothesis that area influences both the number of storks and birthrate, as the correlation pattern we observe among the variables (i.e., the partial regression model described above), can imply more than one underlying causal structure (as already mentioned that is why it is *unresolved*). However, while correlation implies an underlying, unresolved causal structure, *causation always implies a completely resolved correlational structure* (Shipley 2000a). This means that the hypothesized causal relationships among variables imply one, and only one, specific pattern of correlations and partial correlations (which in turn, however, can be cast by more than one causal model). Bill Shipley in his excellent book on cause and correlation in biology compares the pattern of correlations we can observe in nature to the shadows cast on a screen by a three-dimensional object, which in turn represents the causal structure behind the observable correlations (Shipley 2000b). A round shadow can be cast both by a ball as well as by a frisbee (i.e., the implied causal structure is unresolved), but the ball can cast only a round shadow (the implied correlational structure is completely resolved). This means that, at least in principle, we could test the "goodness of fit" of the correlational pattern we would expect to be cast by our hypothesized causal model, with the correlational structure we observe in the data. To be able to do this, we need a formal method to translate between the language of causality and the language of statistical probability. We also need an appropriate measure of the fit between the correlational pattern we observe in the data and the one that must exist given a specific causal structure. We will take a closer look at the recently developed methods permitting us to do exactly this, but first we need to define better the language of causality. To this end, let us complicate a bit our model of the causal relationships linking the various variables present in the Matthews (2000) dataset. For example, we can plausibly hypothesize that while area is the common cause of the number of stork pairs and human birthrate, this latter variable in turn is the causal parent of the number of inhabitants in each country[7] (Fig. 8.4). The causal model depicted in Fig. 8.4 is what in graph theory is called a directed acyclic graph (DAG). Squares represent variables, which in the language of graph theory are called "vertices." The directed arrows, called "edges," represent the hypothesized causal links. The graph is called "Acyclic" because in this kind of graph, a causal path (i.e., the path you can do following the edges passing from one vertex to the next along the causal model) never returns to the same starting vertex. A vertex in a DAG such as birthrate in Fig. 8.4 can be both a

---

[5] Implying that country size is somehow determined by the number of stork pairs inhabiting that country!

[6] Even though it is not necessarily more implausible than the hypothesis that storks deliver babies!

[7] In the data frame storks.dat this variable is called "Humans" and it is expressed as millions of inhabitants.
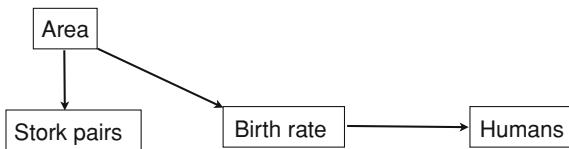
Fig. 8.4 Causal model of the relationship between the number of breeding pairs of storks, human birthrates in European countries, country area and population size depicted as a directed acyclic graph

dependent and an independent variable at the same time (in the language of graph theory you would say that it is both a causal parent and a causal child). We refer to Shipley (2000b) and Pearl (2009) for more details about the language of graph models. DAGs are the mathematical tool we use to formulate hypothesized models of causal relationships among variables. What we now need are formal methods to translate between the language of causality (which we represent with DAGs) and the language of statistical probability. These tools have been introduced to biologists only relatively recently and they go under the name of path analysis and structural equation models (of which classical path analysis is a special case). In the next sections, we will describe them in more detail, with a specific reference to past attempts in the literature to use these methods with data in which data points are represented by species with non-independent errors due to the underlying phylogeny. We will also introduce the *d*-separation-based technique for path analysis (Pearl 1988, 2009) and the *d*-sep test developed by Shipley (2000a), which are at the core of the method we recommend to use for phylogenetic path analysis (von Hardenberg and Gonzalez-Voyer 2013).

8.3 Structural Equations and *d*-Separation-Based Techniques

In structural equation models (SEM), the causal models are translated into a set of linear equations following the causal structure, and the parameters to be estimated from the data are specified. The expected pattern of covariance among the variables can thus be derived simply using the rules of covariance algebra. The free parameters are estimated by maximum likelihood minimizing the difference between the expected covariance matrix of the assumed model and the observed covariance in the data. Finally, we can calculate the probability that the minimum difference between the expected and observed covariance is different from zero (i.e., the observed covariance pattern deviates significantly from the covariance expected by the causal model). This method is appealing because it is based on maximum likelihood and it permits the inclusion of unmeasured latent variables. The latter is an important difference between SEM and path analysis based on *d*-separation, which cannot include latent variables. For a thorough review of SEM

methods, we point readers to Shipley (2000b) and Kline (2010). However, to make SEM methods amenable to work with an underlying phylogenetic signal not only must we compare the covariance matrix expected by the causal model with the covariance observed in the data, but also somehow include the expected covariance due to common ancestry. In Sect. 5, we review past attempts to develop phylogenetic SEMs (for examples, see Lesku et al. 2006; Santos 2009, 2012; Santos and Cannatella 2011). The method that we propose to use for phylogenetic path analysis (von Hardenberg and Gonzalez-Voyer 2013) follows a different approach and is based on the concept of *d*-separation developed by Judea Pearl and his collaborators (Geiger et al. 1990; Pearl 1988; Verma and Pearl 1988).

  D-separation[8] is the 'missing link' between the language of causality, represented as directed acyclic graphs, and the language of statistical linear models. *D*-separation specifies the minimum set (called the basis set) of independence and conditional independence relationships (called *d*-separation statements) that hold true among all variables (the vertices) of the hypothesized causal model. In other words, it specifies the list of all, and only those, pairs of variables that are statistically independent conditioning on a set of other variables in the causal model. The minimum set of conditional independencies is determined in the following manner. First, list all pairs of non-adjacent vertices, i.e., the pairs of vertices that are *not* directly connected by an arrow (edge) in the directed acyclic graph. This gives a list of conditionally independent pairs of variables (these vertices are said to be *d*-separated). Second, list all the vertices with an arrow pointing directly to any of the conditionally independent variables in each pair, i.e., the causal parents of any of the two *d*-separated vertices. This gives the list of variables upon which the independent pairs of variables are conditioned, i.e., the variables that are statistically controlled to test the independence between the *d*-separated variables. Simply combining the two lists, we obtain the minimum set of conditional independence statements, which have to be true not to reject the hypothesized causal model. The conditional independence statements can be directly translated into statistical models using correlation, linear models, or other statistical tests that adequately fit the error structure of the data including nonparametric tests and permutation methods. The flexibility in the statistical methods that can be employed to test the conditional independencies is one of the important advantages of *d*-separation compared to SEM methods. To make the above clearer, we will go back to our "baby-delivering storks" example and the hypothesized causal model depicted in Fig. 8.4. In this simple example, the number of stork pairs is *d*-separated from birthrate (storks, birth), and from human population size (storks, humans). Furthermore, area is *d*-separated from human population size (area, humans). This gives us the following list: [(storks, birth), (storks, humans), and (area, humans)]. Let us now list the causal parents. For the first statement (storks, birth) we have only area, which is directly linked with both vertices. Following the notation proposed by Shipley (2004), we put the parent variables between curled

---

[8] D-separation is an acronym for "Directed" separation.

brackets, in this case: {Area}. For the second statement (storks, humans), we have two parent variables: area, directly causing storks, and birth, which is the causal parent of humans {Area, Birth}. For the last statement in this example (area, humans), we have only one causal parent which is birth directly causing humans {Birth}. The resulting list is [{Area}, {Area, Birth}, {Birth}]. As we mentioned above, combining these two lists we obtain the basis set of conditional independencies which must be true for the data to fit this model: [(Storks, Birth){Area}, (Storks, Humans){Area, Birth}, (Area, Humans){Birth}]. We can now translate these *d*-separation statements to statistical linear models in which we test the independence of the pairs of variables in round brackets conditioning on their parent vertices enclosed in curled brackets. The linear models we get are the following:

Birth $\sim$ Area + Storks
Humans $\sim$ Area + Birth + Storks
Humans $\sim$ Birth + Area

In the OPM (available at: http://www.mpcm-evolution.com), we show how you can test these models using *R*. You may have noticed that actually, we already tested the first of these linear models in Sect. 2 and found that, indeed, human birthrate is statistically independent from the number of stork pairs when conditioning on area with a *p value* of 0.307. The effect of storks on human population size is not significant when conditioning on area and birthrate ($p = 0.110$) as well as the effect of area on humans when conditioning on birthrate ($p = 0.6232$). The fact that none of the three hypotheses implied in the above independence statements is rejected, permits us to say that the hypothesized causal model depicted in Fig. 8.4 is a plausible explanation of the correlation patterns we observe among the variables.

Shipley (2000b) proposed to combine the *p* values using Fisher's C statistic which is calculated with the following formula:

$$C = -2 \sum_{i=1}^{k} (\ln(p_i)) \tag{8.1}$$

where *k* is the number of conditional independencies in the minimum set and *p* their *p* value. The *C* statistic follows a χ^2 distribution with degrees of freedom (*df*)= 2 k. The *C* statistic therefore provides a convenient statistic for testing the goodness of fit of the whole path model. With this test (called the *d*-sep test), the path model is rejected, i.e., it does not provide a good fit to the data, if the p value of the *C* statistic is below the pre-specified alpha value (e.g., 0.05). We can now test the fit of our hypothesized causal model of the relationships among number of stork pairs, human birthrate and population size and country surface area. The C statistic has a value of 7.713, which, knowing that the number of conditional independencies *k* is 3, leads to a *p* value of the *d*-separation test of 0.26. This *p* value is larger than an alpha value of 0.05, and therefore, we can accept the model depicted in Fig. 8.4 as a plausible causal explanation of the
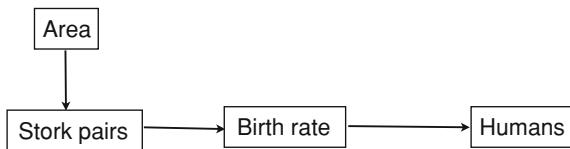
Fig. 8.5 Alternative causal model of the relationship between the number of breeding pairs of storks, human birthrates in European countries, country area, and population size depicted as a directed acyclic graph

relationships found among the variables in our dataset. If you are not convinced, and still believe that storks deliver babies, you can try an alternative model in which instead of having a direct causal link from area to birthrate, you have a direct causal link from the number of stork pairs to birthrate, while the other relationships stay the same as in the previous model. This alternative causal model is depicted as a DAG in Fig. 8.5. We leave it as a little exercise for the readers to obtain the minimum set and thus apply the *d*-sep test to the derived conditional independencies.[9] If you carefully followed all the steps, you should get a C value of 29.2 and a corresponding p value of 5.570647×10^{-5}, which is way below the alpha value of 0.05. This model is therefore rejected, and this should, we think, put the final word on the dispute of whether storks actually deliver babies! In the next section, we show how to apply this elegant and powerful method to data with an underlying phylogenetic signal, introducing in this way our proposed method for phylogenetic path analysis (von Hardenberg and Gonzalez-Voyer 2013).

8.4 A Step-by-Step Guide to Phylogenetic Path Analysis Using the *d*-Separation Method

The first step for any phylogenetic path analysis, as for any study in evolutionary biology, is to clearly define the hypothesis (or hypotheses) being tested. Although this may seem rather trivial to most readers, if not enough time is dedicated to clearly define the hypotheses to be tested, their predictions and underlying assumptions, the study can rapidly go astray and valuable time go to waste. A clear description of the hypotheses to be tested will be crucial for the next step: data collection. Although in the past, the limiting factor for comparative analyses was the lack of well-defined and robust phylogenies, at present the limitations are generally due to insufficient data. A well-defined hypothesis is important to guide researchers as to the data required to test it. We should stress the importance of careful data collection with particular attention to the importance of repeatability,

---

[9] All conditional independencies and full results for this model are provided in the online practical material (http://www.mpcm-evolution.com).

data that are representative of the species, and at the same time are also comparable across species (see Chap. 7).

The second step is to use graph theory to depict the hypotheses being tested as directed acyclic graphs. As mentioned previously, although path analysis is an extension of linear regression, it relies on path diagrams to depict the causal relationships between the variables. Because path analysis is a model-testing procedure, and not a model-developing one, all models to be tested should be based on theory and previous evidence. Once the hypotheses to be compared have been properly depicted by the directed acyclic graphs, the third step is to test the fit of each path model to the data.

As seen above, to test the fit of a path model to the data, we must first enumerate the minimum set of conditional independencies that must be true for the model to adequately fit the data. These conditional independencies can then be translated into linear models and tested with conventional statistical tests. Shipley (2009) showed that the *d*-separation method for path analysis can be extended to data with a hierarchical structure using generalized mixed models to test the conditional independencies in the minimum set. In von Hardenberg and Gonzalez-Voyer (2013), we extended the method further to include the particular case of interspecific comparisons, in which the lack of independence of data points and resulting correlation structure in the residuals violates assumptions of traditional statistical methods. We showed how the conditional independencies can be simply translated into phylogenetic generalized least squares models. Because the conditional independencies are being tested using linear regression models (PGLS) rather than correlations, the order in which we put the variables in the model is important and thus care must be taken when determining which vertex is the *"dependent variable"* and which vertex is the *"independent variable."* Vertices that are causal children (at the end of the causal path separating the two vertices of interest) are *dependent variables*, while causal parents (at the beginning of the causal path) are the *independent variables*. To calculate the number of conditional independencies in the minimum set, the following handy formula can be employed (Shipley 2000b):

$$\frac{V!}{2 \times (V-2)!} - A \qquad (8.2)$$

where V is the number of vertices in the directed acyclic graph and A is the number of edges (the arrows in our DAG). The test, for each conditional independency, involves determining whether indeed, vertices are uncorrelated when conditioning on the parents of each of them. A slightly special case, for defining conditional independencies, is when two vertices are separated by a collider vertex. A vertex is called a collider vertex when two edges from opposite directions in the causal path point toward it (e.g., $A \rightarrow B \leftarrow C$). A collider vertex is said to switch the causal path from *active* to *inactive*, that is, vertices in one side of it are unaffected by changes in vertices in the other side. Hence, when testing the conditional independency of vertices on either side of a collider, the collider is not included in the

conditioning set. For example, to test the conditional independency of (A, C) the conditioning set is $\{\phi\}$. The symbol ϕ is used to indicate that A, C are conditioned on no other variable. The last step is to combine the p values of the conditional independence tests using Fisher's C statistic and thus test the fit of the hypothesized causal model as shown in the previous section (Sect. 8.3).

To illustrate more clearly and in a greater extent the process of phylogenetic path analysis, we now present an empirical example, which we invite the readers to follow replicating it on their own computers using the R language. To make the example biologically meaningful and more intuitive, we propose the following evolutionary puzzle. The aim of the study is to identify the factors that determine geographical range size in the Rhinogrades. If you are unaware of this particular mammal order (Rhinogradentia, DE BURLAS Y TONTERIAS 1948), this comes as no surprise. The Rhinogrades (also called snouters) were endemic to the islands of the Hi-yi-yi Archipelago in the Pacific Ocean only discovered in 1941 but erroneously completely destroyed by secret nuclear experiments in the 1950s, causing the complete extinction of this highly diversified taxonomic group. The main characteristic of the Rhinogrades is that their noses evolved and diverged (in an analogous way to the beak in Darwin finches) into variegated forms with the most diverse functions. In particular, in most genera of the Rhinogradentia, the nose evolved into a complex locomotion organ (Fig. 8.6). For a full account of the natural history of the Rhinogrades, we refer the readers to Stümpke (1967). Previous studies suggest that, as in other mammalian species, there is an allometric relationship between range size and body mass. Some Rhinogradentia specialists suggest that species with larger range sizes also have larger litter sizes, because of higher resource availability. Nonetheless, given the allometric relationship between body size and litter size, it is still unclear whether the association between litter size and range size is causal or merely correlational. There is much discussion regarding the relationship between range size and nose length. On the one hand, range size has been proposed to directly affect nose length, given that Rhinogradentia use their nasal appendage for locomotion and hence larger range sizes select for longer-distance displacements. Alternatively, some experts suggest that the direction of causality should be turned upside down and that it is nose length that determines displacement distances, and therefore, species with longer noses are able to expand their range size. Finally, there is some consensus among Rhinograd experts that dispersal distance is determined by nose length. Based on this knowledge, we can construct a set of hypotheses of causal relationships among variables which we can depict using directed acyclic graphs with the five traits of interest. We will refer to the five traits of this example with acronyms for brevity: body mass = BM, litter size = LS, nose length = NL, dispersal distance = DD, and range size = RS. Figure 8.7 presents the models we proposed for the present example.

The nuclear disaster which destroyed the Hi-yi-yi islands, together with the Rhinogrades also brought down the Darwin Institute of Hi-yi-yi, where all the specimens and life history data of this group were conserved (Stümpke 1967). We therefore had no other choice but to resort to simulated data for our example.
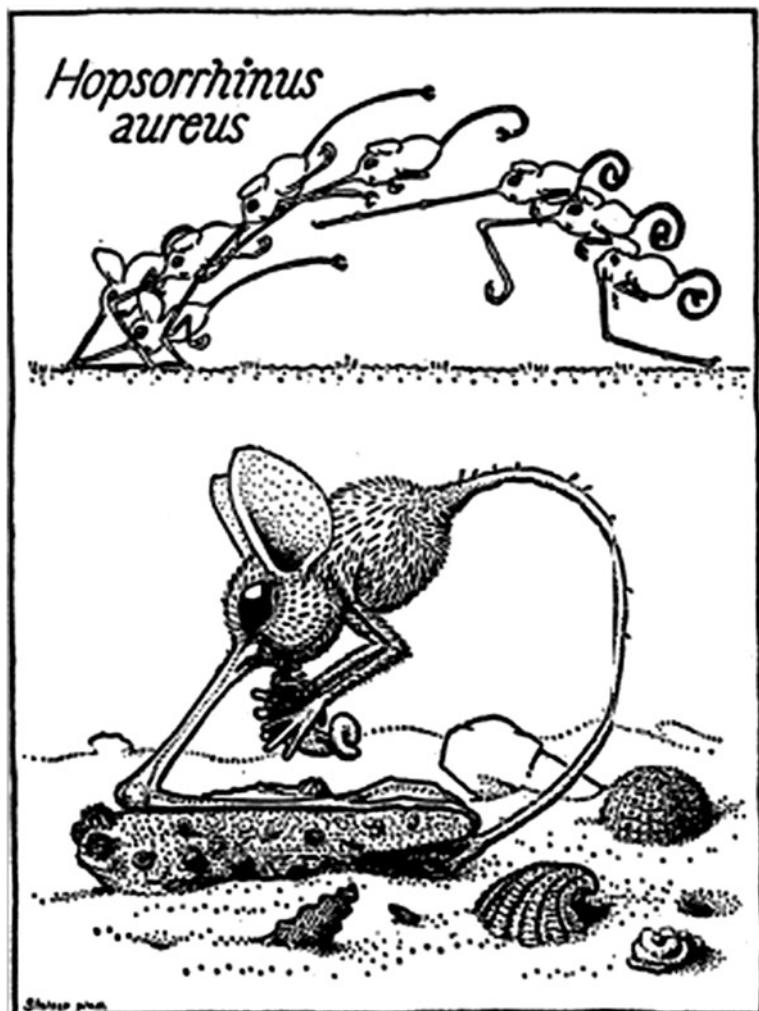
Fig. 8.6 A representative of the Rhinogradentia order: the *Hopsorrinhus aureus* belonging to the Hopsorrhinidae family (Snout Leapers sens. strict.), characterized by the peculiar nasal structure which permits them to move thanks to long backward leaps (Taken from Plate VI, in Stümpke 1967)

We simulated a phylogenetic tree of 100 species under a pure-birth model. Data for the five variables were simulated to evolve on the tree following a Brownian model with a lambda-transformed tree ($\lambda = 0.8$). The data were simulated to evolve with varying degrees of inter-correlation among the variables based on a pre-specified causal model. Variables directly linked in the path model presented correlations of 0.5, while variables with indirect links presented correlations that decreased proportionally with the number of variables separating them, with correlation decreasing by half for each variable in the indirect link (see OPM).
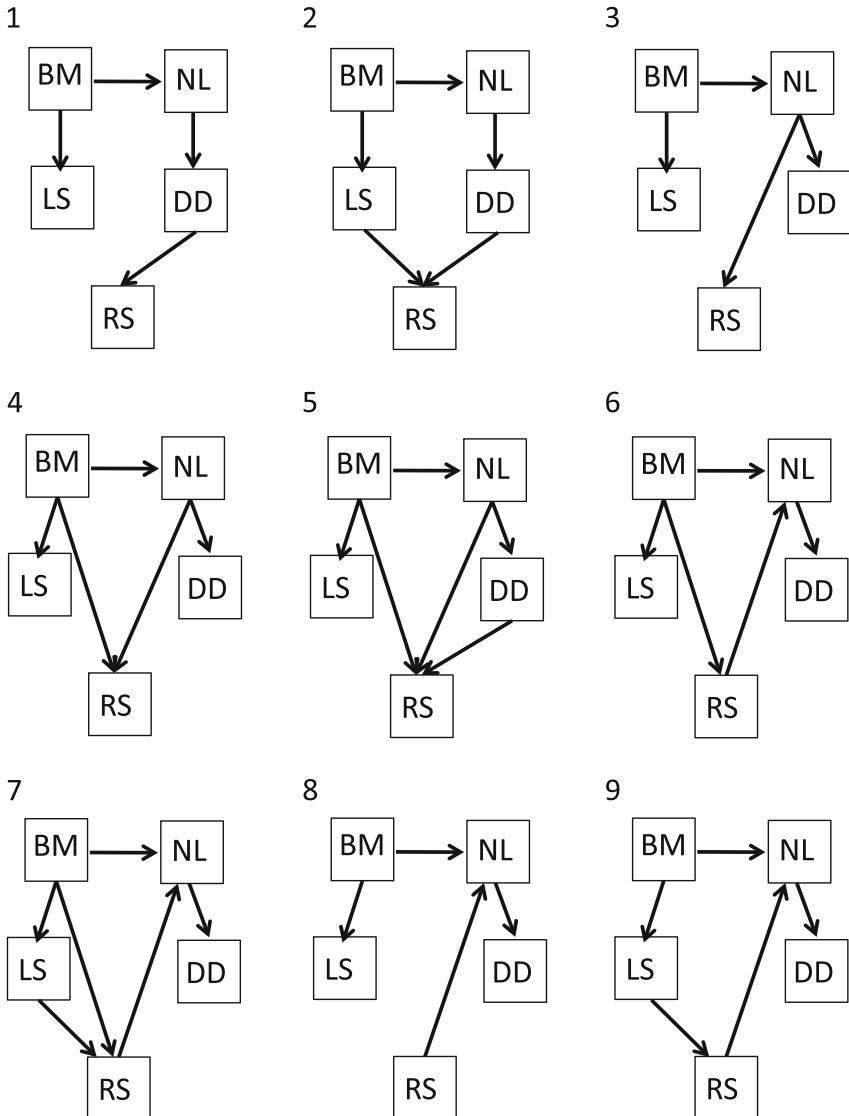
Fig. 8.7 Alternative path models depicting the relationship between body mass, litter size, nose length, dispersal distance, and range size in Rhinogradentia

Use of simulated data following a pre-specified path model is an excellent means to practice a novel approach. In the OPM (http://www.mpcm-evolution.com), we provide both the simulated data (in a file called rhino.csv) and phylogeny (in a file called rhino.tree) used in the Rhinogradentia example, as well as the R code used to simulate the phylogenetic tree and data to enable readers to simulate their own data under a different path model if they wish to do so. We also provide an online tutorial,

replicating all steps described in this section using R. In the first model in Fig. 8.7, there are 5 vertices and 4 edges, hence the minimum set contains 6 conditional independencies to be tested (as follows using the above-mentioned formula).

These are the conditional independencies in the basis set and their translation into linear models:

Conditional independencies	Linear models
(BM, DD) {NL}	DD $\sim$ NL + BM
(BM, RS) {DD}	RS $\sim$ DD + BM
(NL, LS) {BM}	LS $\sim$ BM + NL
(DD, LS) {BM, NL}	LS $\sim$ BM + NL + DD
(LS, RS) {BM, DD}	RS $\sim$ BM + DD + LS
(NL, RS) {BM, DD}	RS $\sim$ BM + DD + NL

It is important to note that since we are using linear models to test the conditional independencies (as opposed to correlations), care must be taken when determining which variable is the response and which the predictor. In such cases, to determine the order of variables in the conditional independency, always follow the direction of causality in the directed acyclic graph (as noted above). Note that in particular cases there is no a priori reason to define one variable as the "response" and the other as the "predictor" as each variable is at the end of a causal path. In such circumstances, the researcher must decide how to define the model to test the conditional independency and keep it constant in other models being compared. We can test the conditional independencies using one of the available statistical packages to perform PGLS and thus obtain the value of the C statistic as described above. Note that an important advantage of the approach we propose is that it allows for analyses to be done using the evolutionary model which best fits the data (Freckleton 2009; Freckleton et al. 2002; Grafen 1989; Hansen 1997; Pagel 1999). In the case of this particular example, given we simulated the data under a Brownian model we will use PGLS analyses with a maximum-likelihood estimate of the lambda parameter (Freckleton et al. 2002; Revell 2010). Following the same steps as for the first path model, we can also test the minimum set of conditional independencies for model 2. These are presented below with their translation into linear models:

Conditional independencies	Linear models
(BM, DD) {NL}	DD $\sim$ NL + BM
(BM, RS) {DD, LS}	RS $\sim$ DD + LS + BM
(NL, RS) {DD, LS, BM}	RS $\sim$ DD + LS + BM + NL
(NL, LS) {BM}	LS $\sim$ BM + NL
(DD, LS) {BM, NL}	LS $\sim$ BM + NL + DD

Table 8.1 C statistic, number of conditional independencies tested (k), and p values of the C statistic for the 9 path models depicted in Fig. 8.7

Model	C statistic	k	p value
1	63.809	6	4.52×10^{-9}
2	62.769	5	1.08×10^{-9}
3	28.973	6	0.004
4	6.582	5	0.764
5	5.258	4	0.730
6	6.439	5	0.777
7	6.018	4	0.645
8	7.699	6	0.808
9	7.362	5	0.691

The reader can now test the conditional independencies of the remaining models depicted as DAGs in Fig. 8.7.[10] Table 8.1 presents the values of the C statistic for each model as well as the p value.

Based on the results in Table 8.1, we can already conclude that the first three models provide very poor fit to the data, because the p value of the C statistic is significant. Hence, we can reject the hypothesis that the correlation structure observed in the data is the result of these three proposed causal models. On the contrary, models 4–9 cannot be rejected, or in other words the correlation structure observed in the data could potentially result from any of these 6 models. As stated above, there is inevitably some uncertainty regarding the causal model that gives rise to the observed correlation structure in the data, and in this case we have identified 6 candidate causal models. This not very satisfactory! Shipley (2000b) proposed two competing models that can be compared based on the difference in the C statistics, which follows a χ^2 distribution with $\Delta \, df = df_{\text{model1}} - df_{\text{model2}}$. However, only nested models can be compared in this manner. Ideally, we would like to be able to compare among all models (including non-nested ones) and rank them based on some estimate of their goodness of fit (Burnham and Anderson 2002). In von Hardenberg and Gonzalez-Voyer (2013), we proposed to use an information theoretic approach alike to the classical Akaike Information Criterion (Akaike 1974) using a modified version of AIC, which we called the C statistic Information Criterion (CIC). This approach was first proposed, in the framework of non-phylogenetic path analysis, by Cardon et al. (2011). Use of an information theoretic approach requires that the measure of "goodness of fit" be based on maximum-likelihood estimates; hence to be able to apply such an approach to path analysis using d-separation, it is necessary to show that the C statistic, used to calculate this criterion, is equivalent to a maximum-likelihood estimate. Shipley (2013) recently provided such mathematical proof, validating the use of AIC (i.e.,

---

[10] All conditional independencies and full results for these models are provided in the online practical material.

CIC) to compare between non-nested models in the framework of d-separation path analysis. This should allay concerns of readers worried by the fact that the C statistic is calculated based on the p values of the conditional independency tests and we are now using it to estimate CICc, apparently combining frequentist and information theoretic approaches. To calculate CIC, we simply need to know the number of parameters estimated in the path model using the empirical data. In phylogenetic path analysis, we assume a multivariate normal distribution of errors and linear relationships between variables, because these are assumptions of the phylogenetic generalized least squares models used to test the conditional independencies (for use of CIC with models with different error distributions, see Shipley 2013). We employ here the formula to calculate CICc, the equivalent of CIC with a correction for small sample sizes. In any case, when the sample size is large relative to the number of parameters, CICc will converge on CIC. To calculate CICc:

$$\text{CICc} = C + 2q \times \frac{n}{(n - 1 - q)} \tag{8.3}$$

where C is the C statistic for the particular model, q is the number of parameters estimated in the path model, and n is the sample size, in the case of phylogenetic path analysis the number of species. For a given path model, we are interested in the slopes of each of the causal links between the variables and the variances. For example, for model 1 in the Rhinogrades exercise, 9 parameters are estimated: the variance for body mass (BM), which is the only variable without any causal parent in the model, and the 4 slopes and the variances for the causal links. While in model 2, 10 parameters are estimated: the variance for body mass, 5 slopes for the causal links and 4 variances, because in this case range size is causally determined by both litter size and dispersal distance, and therefore, two slopes and one variance are estimated for these causal links (see Shipley 2013 for details). In cases in which the interest lies only in the slopes of the causal links between the variables, a quick way to obtain the number of parameters estimated in the model is simply to add the number of vertices and number of edges in the path model. Note that for models to be comparable using CICc, all models must have the same sample size (number of species), and therefore, the data set is reduced to the maximum number of species for which data for all variables is available. Furthermore, all compared models must also have the same number of vertices, although they can have different numbers of edges. Hence, to compare two models in which one variable has no causal link to any other (i.e., there is no edge between it and any other vertex in the model), the complete set of conditional independencies between this variable and all others in the model must be tested to calculate the C statistic. Indeed, such a model assumes that the "isolated" variable (unconnected to any other variable in the model) is conditionally independent from all the variables in the model, and this assumption must be tested (Cardon et al. 2011).

Now we can calculate CICc values for all the models, we are comparing in the Rhinogrades example. With the CICc values in hand, we can also rank the models

based on the difference in CICc (ΔCICc). ΔCICc is simply the CICc value of model i minus the value of the model with the lowest CICc (CICc$_{MIN}$). ΔCICc can be used in the very same way as it is normally done in standard model selection procedures using AIC (Cardon et al. 2011). Given that ΔCICc is measured in a continuous scale of *information,* the values are comparable among models. As a general rule of thumb, models with ΔCICc values < 2 are all considered to have substantial support (Burnham and Anderson 2002). The relative likelihood of a model i given the data L (g$_i$|data), provides information regarding the relative strength of evidence for a model compared to the others and can be computed, following Burnham et al. (2011):

$$\ell = L(g_i|\text{data}) = (\exp -(1/2)\Delta\text{CICc}_i) \tag{8.4}$$

Finally, CICc weights, the probability of each path model g$_i$, given the data and the set of models being compared, are also simple to compute as a measure of strength of evidence (Burnham et al. 2011):

$$w_i = \Pr\{\text{mod}(g_i)|\text{data}\} = \frac{l_i}{\sum_{j=1}^{R} l} \tag{8.5}$$

Use of CICc allows us to move from a hypothesis testing to a hypothesis comparison framework. Below we present the CICc values for all the tested models in the Rhinogrades example, including the number of estimated parameters in each model (q), ΔCICc, likelihoods and weights (Table 8.2).

Use of CICc allows for finer comparisons among models compared with what can be gained by simply looking at the C statistic and its associated p value. Table 8.2 presents a clear ranking of all models from the Rhinogrades example. Those with significant C statistics (models 1, 2, and 3) also show elevated CICc and ΔCICc values, indicating that they provide a very poor fit to the data. We can also however gain some insight about the six models with nonsignificant C statistics. Models 5, 7, and 9 provide a relatively poorer fit to the data than the other three models (4, 6, and 8) as the ΔCICc values are > 2. We cannot distinguish between models 4, 6, and 8, since they present very small differences in CICc, with all models showing ΔCICc values < 2. Note that care must be taken when comparing models with ΔCICc values < 2. As pointed out by Arnold (2010), also highlighted by Burnham and Anderson (2002: p. 131), for equivalent AIC[11] values, care must be taken when interpreting models based solely on ΔAIC (or ΔCICc) values. In some cases, models might not be truly "competitive" with top-ranking models, but appear to be based solely on low CICc values, because addition of an uninformative variable, or in the particular case of path analysis an uninformative causal link between variables, can

---

[11] CICc in the case of phylogenetic path analysis.

Table 8.2 Number of parameters estimated in each model (q) C statistic information criterion with correction for small sample sizes (CICc), ΔCICc, likelihoods (l_i), and CICc weights (ω_i) are shown for each model of the Rhinogradentia example

Model	q	CICc	ΔCICc	l_i	ω_i
8	9	27.700	0.000	1.000	0.349
6	10	28.911	1.211	0.546	0.190
4	10	29.054	1.354	0.508	0.177
9	10	29.834	2.134	0.344	0.120
5	11	30.258	2.558	0.278	0.097
7	11	31.018	3.318	0.190	0.066
3	9	48.973	21.273	2.402×10^{-05}	8.380×10^{-06}
1	9	83.809	56.109	6.548×10^{-13}	2.284×10^{-13}
2	10	85.240	57.540	3.201×10^{-13}	1.117×10^{-13}

lead to marginal changes in CICc values even though there is very little difference in the goodness of fit. Therefore, models with such uninformative causal links might present ΔCICc ≤ 2, generally interpreted as indicating "substantial level of empirical support" (Burnham and Anderson 2002: 170), although such an interpretation would be erroneous. Burnham and Anderson (2002) suggest that models having Δi [ΔCICc] within 0–2 values of the best model should be examined to check whether they differ from the best model by having 1 more parameter and also present essentially the same maximized log-likelihood value (in this particular case, similar C statistic). In such cases, the model with more parameters is not really supported, but presents marginal difference with the "best model" simply because one parameter is added to the model, although the fit of the model is not truly improved as measured by the log-likelihood value (C statistic). Returning to our example, we can see that models 4 and 6 differ by a single parameter from model 8, the best-fitting model. Model 4 also differs from model 8 in the direction of the causal link between range size and nose length, which as the reader might remember was the cause of much discussion among Rhinograd experts. These models also present small differences in C statistic with model 8 (model 4: difference = 1.35, model 6: difference = 1.21). Hence, following Burnham and Anderson (2002) models 4 and 6 might not be considered as supported and competitive to the same degree as the best-fitting model 8, even though they are within ΔCICc < 2. Note that we are by no means advocating selection of a single model over all others. Rather, following Burnham and Anderson (2002) and Arnold (2010), we highlight the need for caution when comparing models, above all that it should not be done mechanistically simply based on ΔAIC (ΔCICc) values. In applications of phylogenetic path analysis with empirical data, it is highly likely that more than one model will present small (< 2) ΔCICc values. Under such circumstances, conclusions should be drawn based on the set of most likely models.

In our example, Model 8 appears to be the best-fitting model. We can now calculate standardized path coefficients of the causal edges linking the variables

according to this model. Standardized path coefficients are particularly useful because, being standardized, they are comparable with each other, and therefore, we can compare the relative strength of each causal relationship in the model. To calculate them, we must first standardize the original data. To do this, we subtract the trait specific population mean from each value and divide by the standard deviation. In the specific case of the simulated data used in this example (given it is randomly drawn from a multivariate normal distribution with mean 0 and standard deviation of 1), the data are already standardized, therefore this step is not necessary.

We then use the standardized data to calculate the standardized path coefficients using PGLS analyses, following the causal paths in the directed acyclic graph. In the case of model 8, the path coefficients are as follows:

BM→ LS 0.4973 (±0.0893 s.e.)
BM→ NL 0.4614 (±0.0650 s.e.)
RS→ NL 0.5281 (±0.0572 s.e.)
NL→ DD 0.6285 (±0.0800 s.e.)

Had we truly competitive models, one way to account for this "model uncertainty" is model averaging (Burnham and Anderson 2002). In von Hardenberg and Gonzalez-Voyer (2013), we showed how standard model averaging procedures can be applied also in the context of phylogenetic path analysis, averaging the path coefficients of all models with CICc < 2 according to the CICc weights of each model, thus on the relative strength of the models in the averaged set of models.

What have we learned regarding the relationship between range size, nose length, and other traits in Rhinogradentia after employing phylogenetic path analysis to tackle the question? First, based on the best-supported model (ΔCICc ≤ 2), range size appears to be the causal parent of nose length, while litter size does not appear to be causally linked to range size. Moreover, the effect of range size on dispersal distance appears to be indirectly mediated through nose length. In other words, in Rhinograds, dispersal distance appears to be directly determined by nose length. Finally, given this entire example was based on data simulated following a pre-specified path model we can now ask how precise is phylogenetic path analysis in identifying the path model giving rise to the data? Well, quite accurate in fact! The model we used to simulate the data is actually model 8, which is the best-supported model based on CICc. Furthermore, model 9 is identical to model 8 except for the additional causal link between litter size and range size. Despite virtually identical C statistics, there is a difference in CICc of 2.13, which suggests CICc is adequately penalizing this model for the additional parameter. Finally, looking at the standardized path coefficients calculated above, we see that they are all roughly around 0.5, which is not surprising, but reassuring, as the data have been simulated with correlation coefficients of 0.5 for all the pre-specified direct links.

8.5 Phylogenetic Non-Independence of Data Points, Correlated Residuals, and the Problems with Inflated Type I Error

The interest in the present chapter is to apply path analysis to macroevolutionary questions, involving comparisons among numerous species. Attempting to convince readers of this book of the importance of accounting for non-independence of data points due to phylogenetic relatedness of species is like preaching to the choir.[12] Nonetheless, we present first the challenges associated with accounting for phylogenetic relatedness in path analysis and second demonstrate the extent of the problem if non-independence of data points is ignored when undertaking confirmatory path analysis using the *d*-separation method (von Hardenberg and Gonzalez-Voyer 2013). It is well known that interspecific comparative analyses violate the assumption of traditional statistical methods that data points are independent, indeed the varying degrees of shared ancestry of the species included in the analysis influences the expected similarity of trait values (Felsenstein 1985; Freckleton et al. 2002; Garland et al. 1992; Harvey and Pagel 1991). For linear models, the main problem is the correlation structure of the residuals that is determined by the degree of phylogenetic relatedness among species (Felsenstein 1985; Grafen 1989; Martins and Hansen 1997; Revell 2010; see Chap. 5). The consequences of not accounting for phylogenetic effects in statistical analyses of multispecies data are, among others, artificially inflated number of degrees of freedom, incorrectly estimated variances, and increased type I error rates of significance tests (Felsenstein 1985; Harvey and Pagel 1991; Martins et al. 2002; Martins and Garland 1991; Rohlf 2006). These problems, however, become compounded in path analysis because of the requirement of testing multiple structural equations (in the case of SEM) or all the conditional probabilistic independencies that must be true for the causal model to be correct (in the case of the *d*-sep test). Previous attempts at controlling for phylogenetic relatedness in path analysis exist. Among those having included an explicit description of how phylogenetic non-independence was controlled are Lesku et al. (2006) and Santos and Cannatella (2011) who used phylogenetic independent contrasts (PIC; Felsenstein 1985) as the data entered in a SEM. Use of independent contrasts allowed the authors to account for phylogenetic non-independence explicitly in their SEM. However, there are limitations associated with the use PIC. First, the method assumes the traits, and covariances between traits evolve following a strict Brownian motion model and performance can be compromised if the assumption is not met (Revell 2010), second, PIC assumes strictly linear relationships between traits (Quader et al. 2004). More recently, Santos (2012) combined two approaches to control for phylogenetic non-independence in SEM in a study aimed at analyzing the factors associated to rate of molecular evolution in poison frogs. First,

---

[12] All pun intended!

for a set of species trait values, he estimated the phylogenetic signal of each trat by estimating the maximum-likelihood value of λ, he then calculated PIC from a λ-transformed phylogeny using the ML estimate for each particular trait. For data on rate of molecular evolution he used an estimate of the variance-covariance matrix derived from a molecular phylogeny.

We proposed an alternative approach (von Hardenberg and Gonzalez-Voyer 2013) combining confirmatory path analysis using the d-separation method (Pearl 1988; Shipley 2000b) and phylogenetic generalized least squares (PGLS; Martins and Hansen 1997). The advantage of PGLS is that it can incorporate distinct models of trait evolution, can combine continuous and categorical variables in a single model without the need to code dummy variables, and provides the value of the y-intercept (Martins and Hansen 1997; see Chap. 5). Further, a key advantage of using PGLS is that it allows for path analyses to be undertaken using taxon-specific trait values rather than contrasts, facilitating interpretation of the results. Finally, in PGLS an evolutionary parameter is estimated simultaneously with model fit, which determines the amount of phylogenetic signal in the data (in the residuals of the model to be precise) and hence the necessary correction for the expected covariance in trait values resulting from phylogenetic relatedness, given the evolutionary model (Freckleton et al. 2002; Martins and Hansen 1997; Revell 2010). This is an important advantage because in some instances data may present a phylogenetic structure that is intermediate between that predicted by the evolutionary model and absence of phylogenetic correlation in the data (Freckleton et al. 2002; Revell 2010). Under such circumstances, PGLS models have been shown to outperform independent contrasts (Martins and Hansen 1997). These advantages of PGLS allow us to ensure that tests of conditional independencies are done with the adequate correction for phylogenetic signal in the residuals of each particular model. Note that the flexibility of the d-separation method also allows researchers to combine continuous, categorical, and discrete variables in their path models, because tests of conditional independencies can be done using phylogenetic ANOVA, or other appropriate statistical methods (see Chap. 12 for an introduction to phylogenetic-mixed models).

In von Hardenberg and Gonzalez-Voyer (2013), we used a simulation-based approach to explore the consequences of ignoring phylogenetic non-independence when undertaking confirmatory path analysis using the d-separation method. We simulated evolution of five hypothetical traits along a simulated phylogeny under the covariance matrix expected from the causal relationships among the traits derived from a specific pre-defined causal model. In order to analyze the effects of varying degrees of phylogenetic signal in the data, the simulations were run under six different scenarios with different degrees of lambda (λ), spanning from null to strong phylogenetic signal in the simulated data. When $\lambda = 0$ traits were simulated evolving along a star phylogeny, where trait evolution for each species is completely independent, while at the other extreme of $\lambda = 1$ traits were simulated to evolve following a pure Brownian motion model, where the degree of similarity between species traits is inversely proportional to the distance to the nearest common ancestor. For the four remaining scenarios, prior to simulating trait
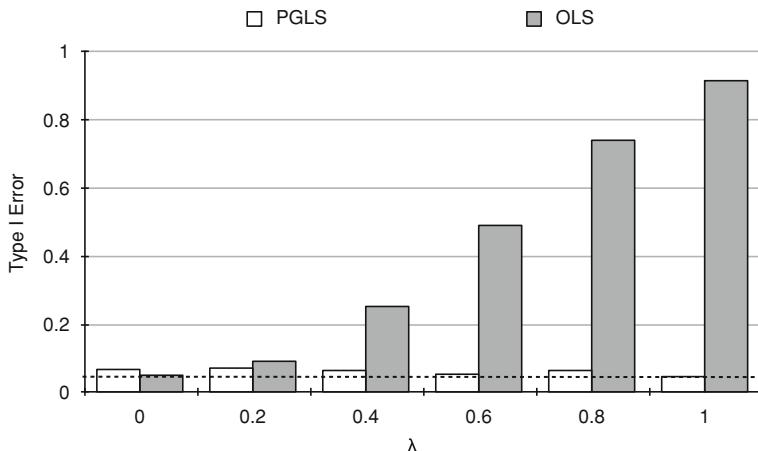
Fig. 8.8 Type I error of traditional (i.e., non-phylogenetic OLS) and phylogenetic (PGLS) path analysis under six simulated scenarios spanning low to high phylogenetic signal in the data

evolution, the phylogenetic tree was transformed based on values of λ ranging from 0.2 to 0.8 (i.e., 0.2, 0.4, 0.6, and 0.8). Tests of conditional independencies were done using the untransformed tree. One thousand datasets were simulated for each of the six scenarios, each with an underlying phylogenetic tree of a fixed, arbitrary size of 100 species. Each simulation of trait evolution was done using a different simulated phylogeny; hence simulations also incorporated the effects of varying phylogenetic topology. At each iteration, von Hardenberg and Gonzalez-Voyer (2013) calculated Fisher's C statistic and obtained a distribution of p values to determine the level of type I error (i.e., the probability of rejecting the null hypothesis, in this case the tested model, when it is true, testing the predicted set of conditional independencies consistent with the "true" underlying causal model) and the power (i.e., 1-the type II error, the probability of not rejecting the tested model when it is actually false, testing the predicted set of conditional independencies of a "wrong" causal model). These simulations were run both for d-sep tests ignoring phylogenetic effects and for the phylogenetically explicit d-sep test. The results of the first test, type I error, are shown in Fig. 8.8. It is clear that the type I error of "classical" path analysis, ignoring phylogenetic non-independence, increases rapidly with the degree of phylogenetic signal in the simulated data to reach values > 0.9 when traits are simulated to evolve via Brownian motion. On the contrary, although our phylogenetic path analysis method is slightly over-conservative, it nonetheless performs well under varying degrees of phylogenetic signal in the data. Figure 8.8 clearly demonstrates the importance (to say the least) of accounting for phylogenetic relatedness when undertaking path analysis using the d-separation method. However, power was in general comparable between "classical" path analysis, ignoring phylogeny, and phylogenetic path analysis (see Fig. 8.9). The high power of non-phylogenetic path analysis is not surprising. The
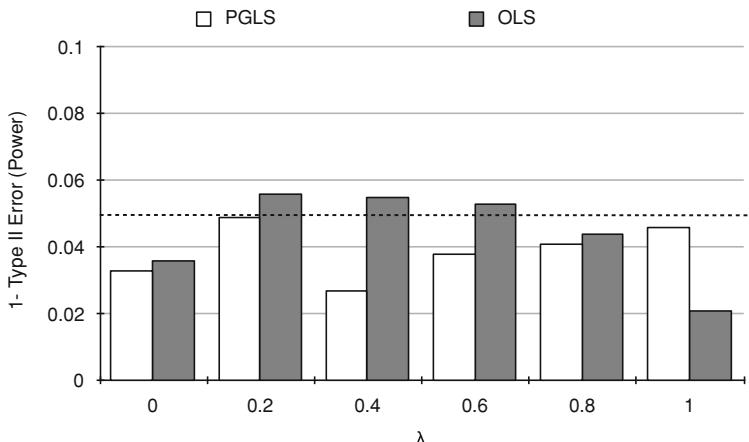
Fig. 8.9 Power of traditional (i.e., non-phylogenetic OLS) and phylogenetic (PGLS) path analysis under six simulated scenarios spanning low- to high-phylogenetic signal in the data

sagacious reader will have already guessed that the high power of non-phylogenetic path analysis is a consequence of the high type I error. Indeed when ignoring phylogenetic relationships, there is a higher probability of detecting significant correlations among traits, even if these are simply due to phylogenetic relatedness rather than true correlated evolution, with the result of a higher probability of rejecting the proposed model.

8.6 Does Collinearity Affect Path Analysis?

Literature on the effect of collinearity on Path Analysis is controversial. While some studies suggest that structural equation models (SEM) can effectively eliminate problems with collinearity (Pugesek and Grace 1998; Pugesek and Tomer 1995), others suggest it can be cause for concern (Petraitis et al. 1996; Grewal et al. 2004). As far as we know, no study has specifically dealt with the effect of multicollinearity on the *d*-separation method. Because the phylogenetic path analysis method we presented (von Hardenberg and Gonzalez-Voyer 2013) is based on the use of PGLS to test conditional independencies, violations of the assumptions of PGLS will inevitably undermine such tests. Least squares estimates of statistical model parameters are robust to moderate, even high, levels of collinearity (Freckleton 2011). However, estimates of parameter variance may be very sensitive affecting hypothesis tests, which would undermine confidence on tests of conditional independencies. Hence, strong collinearity can indeed be a problem, as long it is a problem for PGLS although it will be limited to the specific conditional independences we are testing. Our view is however, that the *d*-separation method can actually be an effective way to disentangle collinearity, at least

when it is not very strong. Indeed, the way you set up your path analysis model, and test for the independence among the variables to see if your model fits the data, you basically are testing for the presence of collinearity among your variables. Models with strong collinearity among the variables not directly causally linked will be rejected by the data and therefore will not be accepted as a possible explanation of the cause–effect relationships among the variables. On the other hand, collinearity between predictors could also affect the power of tests of conditional independencies because collinearity increases the standard error of partial regression coefficients. As collinearity increases, the ability to detect a significant effect (statistically non-zero partial regression slope) is reduced (Freckleton 2011). An often-unappreciated problem is the effect of measurement error, which is common for most (if not all) data employed in comparative analyses. Measurement error can result in underestimation of model parameters, even in the absence of collinearity, due to attenuation (Freckleton 2011). Bias increases when there is measurement error in combination with collinearity. Under such circumstances, attenuation leads to underestimation of the effect of the predictor with the weakest effect, while the predictor with stronger effect is over-estimated (Freckleton 2011). One possibility, which would need to be explored, is to include within-species variation in the models, for example, using mixed models (see Chaps. 7 and 10). By including several measurements per species for each trait, we could not only obtain a better estimate of the species mean but also obtain an estimate of the species-specific variation, which could potentially mitigate the effects of measurement error, although this has yet to be explored in the context of phylogenetic path analysis. We follow Freckleton (2009) and strongly suggest to always verify that the assumptions of the statistical methods employed to test the conditional independencies of the path model are met, this will ensure robust results of tests of conditional independencies.

8.7 Conclusions

The aim of this chapter was first to demonstrate in a didactic and easy to follow manner how to undertake a path analysis using the *d*-separation method (Shipley 2000b), while explicitly accounting for phylogenetic non-independence. As pointed out previously, the method we propose (von Hardenberg and Gonzalez-Voyer 2013) is not the only attempt (see for example Lesku et al. 2006; Santos 2009, 2012; Santos and Cannatella 2011). However, we think our method has some advantages, including, but not limited to, flexibility in the evolutionary model, ability to execute the analysis on the data as such rather than resorting to independent contrasts, and ability to include variables resulting in non-normal distribution of errors. Comparative methods are developing rapidly, for example, Chap. 9 in this book deals with phylogenetic logistic regression methods, which could in theory allow for phylogenetic path analysis including binary traits. Furthermore, the flexibility of PGLS would also allow for phylogenetic path

analysis to be undertaken accounting for variation in species traits (Martins and Hansen 1997), for example, using mixed models. The second aim of this chapter was to show how using phylogenetic path analysis novel questions in macroevolution can be addressed. Using our example with the simulated Rhinogradentia data, we showed how path analysis can help in disentangling evolutionary relationships between traits. For example, based on the results we can say, with some confidence, that litter size has no direct causal effect on range size in this fictitious mammalian order. We also show how phylogenetic path analysis can be employed to compare models with alternative causal relationships between variables. We must once again point out that the observed correlational pattern in the data can imply more than one underlying causal model, hence we might not always be able to distinguish between alternative causal models. Nonetheless, use of CICc, model comparison, and model averaging procedures can allow us to propose causal hypotheses among variables from the observed correlational patterns. Do we mean to say that employing this method we can do away with the limitations of comparative analyses for inferring causality pointed out at the beginning of the chapter? By no means! Such limitations are still there, and the statistical controls we use to disentangle cause–effect relationships are of course not comparable to the physical controls and randomizations we can apply in well-designed experiments. However, as stated at the beginning of the chapter, such an experimental approach is virtually impossible to carry out in the context of comparative analyses. Phylogenetic path analysis (using the *d*-separation method we propose or other approaches) may well be the only resort we have to infer causality in comparative studies. We must however keep in mind that path analysis is a hypothesis testing approach rather than a hypothesis-generating method. Carefully pondered and biologically meaningful, and supported, hypotheses of the causal relationships among studied traits must be presented before jumping into model testing. The end result of such a process is the confirmation of the plausibility of the proposed evolutionary causal model (although other alternative causal models can possibly explain the same observed correlation pattern), and probably more interestingly, the rejection of erroneous evolutionary causal models. We would therefore caution readers against overconfidence on the correctness of a causal model fitting the observed correlation structure; nonetheless we can be reasonably sure that rejected models are wrong. With all the uncertainties macroevolutionary studies must deal with, we think that the advantages provided by phylogenetic path analysis are not trivial. Furthermore, the causal model, or the set of models, we finally adopt as potential evolutionary explanations of the patterns we observe among the traits, can be formally challenged by alternative models in future studies involving new or better data. Such a process of presentation of a model (our causal hypothesis) and its provisory acceptance as plausible explanation of a causal phenomenon until it is confuted by an alternative model is at the very base of modern scientific methodology. We hope to have been successful in transmitting our enthusiasm for this method and to stimulate thought as to how it can allow you to tackle evolutionary questions in the context of comparative analyses.

References

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19(6):716–723

Arnold TW (2010) Uninformative parameters and model selection using Akaike's information criterion. J Wildl Manage 74(6):1175–1178

Burnham KP, Anderson DR (2002) Model selection and multimodel inference: a practical information-theoretic approach. Springer, New York

Burnham KP, Anderson DR, Huyvaert KP (2011) AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. Behav Ecol Sociobiol 65:23–35

Cardon M, Loot G, Grenouillet G, Blanchet S (2011) Host characteristics and environmental factors differentially drive the burden and pathogenicity of an ectoparasite: a multilevel causal analysis. J Anim Ecol 80:657–667

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

Fisher RA (1926) The design of experiments, 1st edn. Oliver and Boyd, Edinburgh

Freckleton RP (2009) The seven deadly sins of comparative analysis. J Evol Biol 22(7):1367–1375. doi:10.1111/j.1420-9101.2009.01757.x

Freckleton RP (2011) Dealing with collinearity in behavioural and ecological data: model averaging and the problems of measurement error. Behav Ecol Sociobiol 65(1):91–101. doi:10.1007/s00265-010-1045-6

Freckleton RP, Harvey PH, Pagel M (2002) Phylogenetic analysis and comparative data: a test and review of evidence. Am Nat 160(6):712–726. doi:10.1086/343873

Garland TJ, Harvey PH, Ives AR (1992) Procedures for the analysis of comparative data using phylogenetically independent contrasts. Syst Biol 41:18–32

Geiger D, Verma T, Pearl J (1990) Identifying independence in Bayesian Networks. Networks 20:507–533

Grafen A (1989) The phylogenetic regression. Phil Trans Roy Soc B 326:119–157

Grewal R, Cote JA, Baumgartner H (2004) Multicollinearity and measurement error in structural equation models: Implications for theory testing. Mark Sci 23(4):519–529

Grim T (2008) A possible role of social activity to explain differences in publication output among ecologists. Oikos 117(4):484–487

Hansen TF (1997) Stabilizing selection and the comparative analysis of adaptation. Evolution 51(5):1341–1351

Harvey PH, Pagel MD (1991) The comparative method in evolutionary biology. Oxford University Press, Oxford

Kline RB (2010) Principles and practice of structural equation modelling methodology in the social sciences, 3rd edn. Guilford Press, New York

Lesku JA, Amlaner CJ, Lima SL (2006) A phylogenetic analysis of sleep architecture in mammals: the integration of anatomy, physiology, and ecology. Am Nat 168(4):441–443

Martins EP (2000) Adaptation and the comparative method. Trends Ecol Evol 15(7):296–299

Martins EP, Diniz-Filho JA, Housworth EA (2002) Adaptation and the comparative method: a computer simulation study. Evolution 56:1–13

Martins EP, Garland T (1991) Phylogenetic analyses of the correlated evolution of continuous characters: a simulation study. Evolution 45(3):534–557

Martins EP, Hansen TF (1997) Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. Am Nat 149(4):646–667

Matthews R (2000) Storks deliver babies ($p = 0.008$). Teach Stat 22(2):36–38

Messerli FH (2012) Chocolate consumption, cognitive function, and Nobel laureates. New Engl J Med 367(16):1562–1564

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

Pagel M, Meade A (2006) Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov Chain Monte Carlo. Am Nat 167(6):808–825

Pearl J (1988) Probabilistic reasoning in intelligent systems. Morgan and Kaufmann, San Mateo

Pearl J (2009) Causality: models, reasoning and inference. Cambridge University Press, Cambridge

Petraitis PS, Dunham AE, Niewiarowski PH (1996) Inferring multiple causality: the limitations of path analysis. Funct Ecol 10:421–431

Pugesek BH, Grace JB (1998) On the utility of path modelling for ecological and evolutionary studies. Funct Ecol 12:853–856

Pugesek BH, Tomer A (1995) Determination of selection gradients using multiple regression versus structural equation models (SEM). Biometrical J 37:449–462

Quader S, Isvaran K, Hale RE, Miner BG, Seavy NE (2004) Nonlinear relationships and phylogenetically independent contrasts. J Evol Biol 17:709–715. doi:10.1111/j.1420-9101.2004.00697.x

Revell LJ (2010) Phylogenetic signal and linear regression on species data. Meth Ecol Evol 1(4):319–329. doi:10.1111/j.2041-210X.2010.00044.x

Rohlf FJ (2006) A comment on phylogenetic correction. Evolution 60(7):1509–1515

Santos JC (2009) The implementation of phylogenetic structural equation modeling for biological data from variance-covariance matrices, phylogenies, and comparative analyses. The University of Texas at Austin, Austin

Santos JC (2012) Fast molecular evolution associated with high active metabolic rates in poison frogs. Mol Biol Evol 29(8):2001–2018

Santos JC, Cannatella DC (2011) Phenotypic integration emerges from aposematism and scale in poison frogs. Proc Natl Acad Sci USA

Shipley B (2000a) A new inferential test for path models based on directed acyclic graphs. Struct Equ Model 7(2):206–218

Shipley B (2000b) Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference. Cambridge University Press, Cambridge

Shipley B (2004) Analysing the allometry of multiple interacting traits. Perspect Plant Ecol Evol Syst 6(235):241

Shipley B (2009) Confirmatory path analysis in a generalized multilevel context. Ecology 90:363–368

Shipley B (2013) The AIC model selection method applied to path analytic models compared using a d-separation test. Ecology 94(3):560–564

Stümpke H (1967) The Snouters: form and life of the Rhinogrades (trans: Doubleday & Company I). University of Chicago Press, Chicago

Team RDC (2013) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Verma T, Pearl J (1988) Causal networks: semantics and expressiveness. In: Schachter R, Levitt TS, Kanal LN (eds) Uncertainty in artificial intelligence, vol 4. Elsevier, Amsterdam, pp 69–76

von Hardenberg A, Gonzalez-Voyer A (2013) Disentangling evolutionary cause-effect relationships with phylogenetic confirmatory path analysis. Evolution 67(2):378–387. doi:10.1111/j.1558-5646.2012.01790.x

Wilkinson GN, Rogers CE (1973) Symbolic description of factorial models for analysis of variance. Appl Stat 22(3):392–399

Chapter 9
Phylogenetic Regression for Binary Dependent Variables

Anthony R. Ives and Theodore Garland Jr.

Abstract We compare three methods for phylogenetic regression analyses designed for binary dependent variables (traits with two discrete states) both with each other and with "standard" methods that either ignore phylogenetic relationships or ignore the binary character of the dependent variable. In simulations designed to reveal statistical problems arising in different methods, PLogReg (Ives and Garland 2010) performed better than PGLMM (Ives and Helmus 2011) and MCMCglmm (Hadfield 2010) to identify phylogenetic signal in the absence of independent variables; PLogReg also outperformed a standard method for detecting phylogenetic signal in binary data, ancestral character estimation (Schluter et al. 1997; Pagel 1994). All three phylogenetic methods performed similarly for identifying relationships with a continuously valued independent variable x, with all methods having at most moderately inflated Type I error rates, and MCMCglmm having slightly greater power. In contrast, standard logistic regression that ignores phylogeny had seriously inflated Type I errors when x had phylogenetic signal. Perhaps surprisingly, phylogenetic regression that ignored the binary nature of the dependent variable, RegOU (Lavin et al. 2008), performed as well or better than the other methods, at least for larger sample sizes (≥ 64 species), although this approach does not result in a model that can be used to simulate data (e.g., for bootstrapping). We also apply the methods to a data set describing whether antelopes fight or flee versus hide from predators as a function of their group size (Brashares et al. 2000). We end with rough guidelines for analyzing binary dependent variables, with the main recommendation being that multiple methods and simulations should be used to give confidence in the statistical results.

A. R. Ives (✉)
UW-Madison, Madison, WI 53706, USA
e-mail: arives@wise.edu

T. Garland Jr.
UC-Riverside, Riverside, CA, USA
e-mail: theodore.garland@ucr.edu

9.1 Introduction

Generally speaking, comparative data from phylogenetically related species (or higher taxa) cannot be analyzed using standard statistical procedures because mean values for a set of species (or their residuals from a statistical model) are unlikely to be independent and identically distributed (Felsenstein 1985; Garland et al. 1992; Harvey and Pagel 1991). Computer simulation studies have shown, for example, that ignoring phylogenetic correlations often leads to Type I errors, rejecting a null hypothesis that is in fact true (Grafen 1989; Martins and Garland 1991; Diaz-Uriarte and Garland 1996; Martins et al. 2002). The resulting falsely low P-values in statistical tests may lead to seriously wrong conclusions in comparative studies. Thus, use of statistical methods that incorporate phylogenetic information is essential. For continuously valued traits, a growing number of methods and associated software have been developed which incorporate the possibility of phylogenetic correlations among taxa.

Traits with non-Gaussian distributions are surprisingly more difficult to analyze than continuously valued traits when they occur as dependent (response) variables. This difficulty arises in part because most non-Gaussian distributions have means and variances that are not separable. In other words, the variance of the distribution depends on the mean. This is different from the familiar Gaussian distribution in which one parameter gives the mean and a second gives the variance. For example, for a binomial distribution with probability p and number of trials n, the mean is np and the variance is $np(1-p)$, so it is not possible to change n and/or p in any way that changes the mean without also changing the variance. This complication can be approximately addressed by transforming data. For example, if binomial data are divided by $\sqrt{n}$, then the variance is $p(1-p)$ which can be changed independently of the mean $\sqrt{n}p$, thus allowing the data to be fit with a standard regression model that assumes a Gaussian distribution of residuals. This is still an approximation, however, because the regression model assumes that the data can take any value, not just integers between 0 and n. Therefore, the model does not really fit the process that underlies the data and hence cannot be used to simulate data with the same statistical properties. An inability to simulate data makes it difficult to determine the performance of statistical methods designed to analyze such data and generally precludes the use of methods based on simulations to obtain null distributions of test statistics. To solve these problems, the last 20 years have seen huge advances in methods designed to analyze non-Gaussian distributions, and these have spread into phylogenetic comparative methods.

One common type of inherently non-Gaussian response variable is binary, such as when an organism either does or does not possess a particular phenotypic trait (e.g., wings). Three methods have been developed for phylogenetically informed analysis of binary dependent variables: (i) phylogenetic logistic regression, implemented in a MATLAB program named PLogReg (Ives and Garland 2010), (ii) generalized linear mixed models with frequentist estimation, PGLMM (Ives and Helmus 2011), and (iii) generalized linear mixed models with Bayesian

estimation, MCMCglmm (Hadfield 2010). These methods address the dual challenges of formulating an appropriate statistical model and estimating parameters from the model. By accounting for the binary nature of the dependent variable, the hope is that these methods maximize statistical power, that is, the ability to identify parameters that differ statistically from zero or some other value of interest. Power is a major concern, because binary variables contain relatively little information as compared to continuously valued variables (Ives and Garland 2010); for binary variables, information is only available in the form of zeros and ones, with no finer gradation between values. Therefore, there is a premium on methods for binary data that have high statistical power.

When these three phylogenetic methods are used without independent variables, they become tests of phylogenetic signal (sensu Blomberg et al. 2003) in the response variable of interest (Chap. 5). Here, we first compare the three methods in their abilities to identify phylogenetic signal, including two additional "standard" methods: (iv) phylogenetic regression ignoring the binary nature of the dependent variable and assuming an Ornstein–Uhlenbeck model of residual trait variation (Chap. 15), RegOU (Lavin et al. 2008) and (v) maximum likelihood (ML) estimation of discrete traits evolving along a phylogenetic tree using ancestral character estimation, ACE (Pagel 1994; Schluter et al. 1997).

We also compare (i) PLogReg, (ii) PGLMM, and (iii) MCMCglmm for the simple regression case of a single continuously valued independent (predictor) variable, focusing on their abilities to estimate and perform statistical tests on a regression coefficient. In this comparison, we additionally include (iv) RegOU and (v) non-phylogenetic logistic regression with a Firth correction, Logistf from **logistf** {R} (Heinze et al. 2013). We include RegOU because it runs quickly in various software implementations and it might perform adequately in many cases, even though it ignores the binary nature of the dependent variable. We include Logistf because it is a standard method and illustrates the mistakes that can be made by not accounting for phylogeny.

Additional methods that address related statistical problems are not included in our comparisons. For example, Pagel (1994) presents a method for estimating the correlation between two phylogenetically related binary traits. We excluded this approach because here we focus on regression rather than correlation. Felsenstein (2012) presents an estimation procedure for threshold models that can test the correlation between both binary and continuously valued dependent variables. When applied to only binary variables, this method should be similar to MCMCglmm (Hadfield and Nakagawa 2010), so we do not include it here. Finally, note that the methods we analyze are for binary dependent variables. Binary independent variables with continuously valued dependent variables present no special problems for the existing phylogenetic methods for regression and ANOVA or ANCOVA (Garland et al. 1993; Lavin et al. 2008; Dlugosz et al. 2013; Revell 2012; Rezende and Diniz 2012).

A subtext to this chapter is that there is no single "correct" way to perform phylogenetic regressions with binary dependent variables. All of the methods have strengths and weaknesses that must be balanced for a specific data set and

question. The lack of a single best method is not uncommon in statistics when confronted with data arising from complicated processes. The best that can be done is to use different methods, know which is likely to perform best under circumstances resembling those of the data under analysis, and use this knowledge to help select the "best" result. We hope this chapter provides a rough guide for doing this. Despite providing a guide, we hope that we also show that there is no substitute for careful analysis of any complicated set of data, including applying multiple methods and simulations to rigorously confirm the results.

Below, we first give descriptions of the methods. We then perform simulations to compare them, first for the case without independent variables where the methods become tests for phylogenetic signal and then for a simple regression with one independent variable. These simulations are by no means exhaustive; we use them more to illustrate the issues that arise in statistical analyses of binary data and possible ways to address these issues, rather than to give recommendations for which method to use for a particular data set and question. Indeed, one of our main points is that there is no best method for all situations, so you should use our simulations as guides to the types of simulations you should perform. Finally, we apply the methods to a comparative data set which was analyzed previously using phylogenetic logistic regression (Ives and Garland 2010). The online practical material (http://www.mpcm-evolution.com) presents this analysis as a tutorial and also the new code in R that performs the PGLMMs.

9.2 Description of Statistical Models and Estimation Procedures

In building a statistical model, it is often valuable to consider the underlying processes that might generate a data set. For example, for continuously valued traits (e.g., body mass or length), a simple Brownian motion (BM) model (a random walk in continuous time, Chap. 5) could be used to model phenotypic changes that occur in a population experiencing no selection but nonetheless evolving because of random mutation and genetic drift (Felsenstein 1985; Freckleton et al. 2002; Blomberg et al. 2003). Equally, however, the BM model could describe species under strong selection that track the environment instantaneously provided the environment itself changes randomly according to a BM process. The logical conclusion from this recognition is that the observed pattern of trait values among extant species does not necessarily give a lot of information about the processes generating this pattern (Revell et al. 2008). Nonetheless, building a statistical model under a specific evolutionary assumption can lead to a model with useful statistical properties. For example, the BM model can be modified to accommodate stabilizing selection using the Ornstein–Uhlenbeck process, borrowed from physics (Felsenstein 1988; Garland et al. 1993; Martins and Hansen 1997; Chap. 15). In this case, varying the strength of stabilizing

selection varies the strength of phylogenetic correlations between species, thereby producing a statistical model that can be used to estimate phylogenetic signal (Martins and Hansen 1997; Hansen 1997; Blomberg et al. 2003).

For binary traits, no single generally applicable evolutionary or statistical model exists. The lack of a uniquely suitable model is common in statistics, at least after graduating beyond the simplest problems. It is not an issue restricted to the realm of phylogenetically informed statistical procedures. For example, for regression with a binary response variable in the absence of phylogenetic considerations, both logistic and probit models are used routinely, each with its own advantages and disadvantages (Gelman and Hill 2007). We now describe the five approaches that we will compare. We give largely heuristic descriptions of the methods, as opposed to technical descriptions, because these methods are already in the literature.

9.2.1 Phylogenetic Logistic Regression (PLogReg)

The overall structure of PLogReg (Ives and Garland 2010) looks like that of standard logistic regression. A single trait Y can take only one of two values (0 or 1) with probability p, which itself depends on the independent (predictor) variable x. Multiple independent variables can be used, and they can be binary, multistate (coded into a set of 0–1 dummy variables), or continuously valued. A formal specification of the model with a single continuously valued independent variable x is

$$\Pr(Y = 1) = p$$
$$\text{logit}(p) = b_0 + b_1 x \qquad (9.1)$$
$$\text{cov}(Y) = \mathbf{V}(p, a)$$

Here, b_0 and b_1 are regression coefficients, and $\text{logit}(p)$ is the logit function $\log(p/(1 - p))$ that maps any value of p in the interval (0, 1) onto values of $b_0 + b_1 x$ between $-\infty$ and ∞. Thus, for a given value x, the probability that $Y = 1$ is $p = \text{logit}^{-1}(b_0 + b_1 x)$. No term is included for the "residual variation" because for a binary stochastic process, the variance is determined by the mean; specifically, the variance equals $p(1 - p)$. Thus, this is different from a conventional least-squares regression equation, in which an explicit vector of residual deviations from the predicted values is an inherent part of the statistical model. Even though the variance of the binary dependent variable is specified by the mean, the anticipated covariances can be positive as specified by the covariance matrix $\mathbf{V}(p, a)$. These covariances represent phylogenetic signal, that is, the lack of independence among data points Y caused by taxa having experienced a shared evolutionary history prior to the speciation event(s) that begat them. Because the variances depend on the mean p, so too do the covariances, and this presents complications in the statistical model building and estimation not only for the

regression coefficients b_0 and b_1 but also for the parameter that gives the strength of phylogenetic signal, which in PLogReg is called a (Ives and Garland 2010). Although Eq. (9.1) is written with a single independent variable x, PLogReg can accommodate multiple independent variables, and when used without any independent variables (with only regression parameter b_0), PLogReg becomes a method to estimate phylogenetic signal given by the parameter a.

The evolutionary model used to build PLogReg assumes that Y evolves along a phylogenetic tree. There is a constant "instantaneous" probability of the trait switching from 0 to 1 or from 1 to 0, so the more time that elapses, the greater the chance of a switch occurring. The branch lengths of the phylogenetic tree scale the time between nodes, so the probability of a switch occurring between nodes increases with the branch length between nodes. Repeated switches can occur, and an even number of switches results in no difference in Y between two nodes (branching points) on the tree.

Phylogenetic signal arises because any two related species will have the same trait value at their nearest shared ancestral node, just before the speciation event. Depending on the switching rate for the trait in question, these two daughter species will be more or less likely to retain that ancestral state. If the switching rate is very low, then both daughters will likely retain the ancestral state, resemble each other, and hence provide evidence of phylogenetic signal. The overall switching rate given by the parameter a measures the strength of phylogenetic signal in trait Y. The parameter a is scaled so that larger values of a correspond to greater phylogenetic signal (lower switching rates). Although mathematically a can take any real value, numerically PLogReg limits values of a to range between -4 (no signal) and 4 (very strong signal).

Even though greater phylogenetic signal occurs for larger a, it may become very difficult to test for this signal statistically (Ives and Garland 2010). For example, in the extreme case of very strong phylogenetic signal (low switching rates), all taxa will likely share the same trait value (0 or 1), so there is no variation with which to test for statistical significance. This leads to the somewhat counterintuitive expectation that the power to detect phylogenetic signal occurs at intermediate strengths of the signal. Again, this is different from continuously valued traits where, despite limiting the divergence between taxa, strong phylogenetic signal can nonetheless be (strongly) detected in what variation in trait values is observed (Revell et al. 2008).

The process model for trait evolution just described does not involve any independent variables. PLogReg introduces an independent variable x after the evolutionary process establishes phylogenetic correlations in the values of Y among taxa. To model the effects of independent variables, starting with the values of Y following evolution up the phylogenetic tree, the model assumes that these values then rapidly evolve toward 0 or 1 depending on the value of x, independently for each species. In other words, the part of evolution of Y which is driven by the value of x does not depend on phylogeny. Although this is not a realistic model for many scenarios describing the evolution of dependent variables,

it makes sense statistically, because it ensures that if there is no phylogenetic signal in Y, then the model degenerates to standard logistic regression; in general, it is desirable to have phylogenetic methods give their conventional counterparts when there is no phylogenetic signal (Blomberg et al. 2003). This is an example of trade-offs that must sometimes be made for statistical necessity: Although it might not be evolutionarily plausible, the model underlying PLogReg leads to useful statistical properties, and the model parameters can be statistically fit.

Although this model can be statistically fit, doing so is not easy for two reasons that make it impossible to use standard statistical approaches and readily available software. First, the likelihood function of the model is complicated. The likelihood function is central to parameter estimation; it gives the likelihood of observing the data, given values of the model parameters. Therefore, the maximum likelihood (ML) parameter estimates are computed as those that give the greatest likelihood. For parameter estimation in PLogReg, we used the statistical approach of quasi-likelihood functions, which for technical reasons are well suited for logistic regression (see Ives and Garland 2010). The second statistical issue is that ML estimation for standard (non-phylogenetic) logistic regression is biased; standard ML estimates are on average further from zero than they should be (Heinze and Schemper 2002). In the non-phylogenetic case, this bias can be largely corrected by penalizing the likelihood function as suggested by Firth (1993). We used a similar approach in PLogReg. Simulations showed that this improves the statistical properties of the estimates of PLogReg parameters (Ives and Garland 2010).

Simple diagnostics for determining the adequacy of models for binary dependent variables do not exist, especially for the phylogenetic case. Of course, it is always important to plot the data and fitted model, which can be instructive for identifying gross violations of model assumptions. Nonetheless, the best approach to assess the quality of parameter estimates is to perform a "parametric bootstrap" (Efron and Tibshirani 1993). A parametric bootstrap takes a fitted model and uses it to simulate (a large number of) data sets. The parameters are then reestimated for each of the simulated data sets. Some deep statistical theory shows that the distribution of the parameter values estimated from the simulated data sets approximates (i.e., approaches asymptotically) the theoretical distribution of the parameter estimates (Efron and Tibshirani 1993). This distribution can then be used to obtain confidence intervals and perform statistical tests regarding the parameters. For PLogReg, it is possible to approximate the distributions and confidence intervals of estimates of the regression parameters b_0 and b_1, but bootstrapping is the only effective way to obtain this information for the phylogenetic signal parameter a (Ives and Garland 2010).

It is also possible to perform a bootstrap to test null hypotheses, for example, that a regression coefficient is zero, $H_0: b_1 = 0$. This is done by fitting the data assuming $b_1 = 0$, simulating the model to produce bootstrap data sets, estimating b_1 for each simulated data set, and counting the number of values of b_1 for the simulated data sets that exceed the value of b_1 calculated from the data. In practice, however, this approach often gives very similar results to that of bootstrapping

using the observed value b_1 and testing $H_0:b_1 = 0$ using the bootstrapped confidence intervals of b_1.

An additional advantage of parametric bootstrapping is that it allows identification of bias in parameter estimates. If, for example, the mean of the parameter estimates from the simulated data sets is lower than the value obtained from fitting to the real data and consequently used to perform the simulations, then this would indicate downward bias in the estimates, including the estimate from the original data. It is possible to use this information to correct for bias; a value is picked that, when used in the simulations, produces a mean parameter value that matches the value computed from the data. However, we do not pursue this form of bootstrap bias correction here, instead using bootstrapping simply to identify the existence of bias. Although bootstrapping is useful, it is not a panacea, and there is no substitute for looking at the data and fitted model, and using different methods.

In the simulations, we used the MATLAB code for PLogReg (Ives and Garland 2010), although there is a fast version available in **phylolm** {R} (Ho and Ane 2014).

9.2.2 Generalized Mixed Model with Frequentist Estimation (PGLMM)

PGLMM (Ives and Garland 2010) is a "phylogenetic" implementation of a generalized linear mixed model (Gelman and Hill 2007; McCulloch et al. 2008; Bolker et al. 2009) for binary data. The PGLMM for a single independent variable x is

$$
\begin{aligned}
\Pr(Y = 1) &= p \\
\mathrm{logit}(p) &= b_0 + b_1 x + \varepsilon \\
\boldsymbol{\varepsilon} &\sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{C})
\end{aligned}
\tag{9.2}
$$

Unlike PLogReg, PGLMM treats the probabilities p as random variables, with the distribution of $\mathrm{logit}(p)$ being given by a standard regression model that includes the random variable ε that contains phylogenetic information. The value of ε can be thought of as a continuously valued, phylogenetically inherited, but unmeasured trait. We assume that it evolves up the phylogenetic tree according to a BM evolutionary process (Chap. 5). This makes ε a Gaussian random variable with covariance matrix $\sigma^2 \mathbf{C}$ in which diagonal elements c_{ii} are proportional to the branch lengths from the basal node of the phylogenetic tree to taxa i, and off-diagonal elements c_{ij} are proportional to the shared branch lengths between taxa i and j. The diagonal elements of $\mathbf{C}$ can be equal (for an ultrametric tree with contemporaneous tips) or unequal (e.g., for a tree with time-calibrated branch lengths in which some species became extinct in the distant past).

An evolutionary interpretation of PGLMM is that an underlying, unobserved continuously valued trait evolves up a phylogenetic tree. Then, for taxa at the tips of the tree, the value of this trait determines the probability of Y taking values 0 or 1. Therefore, two stochastic processes are in play: the evolution of the underlying continuous trait that gives the value of p and the choice of the value of Y given probability p. For example, a herbivorous insect might evolve increased production of a detoxifying enzyme (a continuous trait) that increases the chances that it can shift to use a new host plant containing high levels of the toxin (a binary trait: adoption or not of the host plant). Alternatively, the expression level of some gene might determine whether wings develop during ontogeny. The PGLMM model differs slightly from "threshold" models (Felsenstein 1988, 2012) in which the value of Y is determined strictly according to whether the underlying continuous trait exceeds a threshold value, although threshold and PGLMM models are broadly equivalent from a functional perspective (Hadfield and Nakagawa 2010).

In PGLMM, σ^2 measures phylogenetic signal. If $\sigma^2 = 0$, Eq. (9.2) reduces to standard logistic regression (McCullagh and Nelder 1989). If $\sigma^2 > 0$, then the phylogenetic covariances between values of ε lead to covariances between the values of Y. Unlike PLogReg, there are no covariances between values of Y other than those contained in $\mathbf{C}$ and generated by ε. A important statistical property of this model is that the variances (diagonal elements of $\mathbf{C}$) are redundant. For example, assume that the phylogenetic tree is a "star" with unconnected branches of equal length leading from basal node to each tip (terminal) taxon. The corresponding phylogenetic covariance matrix has identical diagonal elements and zero off-diagonal elements. Therefore, even if $\sigma^2 > 0$, no phylogenetic signal exists. As a consequence, the variances have no real effect in the model: The value of σ^2 does not affect the variance in the values of Y, because we know that the variance equals $p(1 - p)$. The only effect σ^2 has is to complicate the interpretation of the regression coefficients b_0 and b_1. Because the logit function is nonlinear, the expected value of p is not equal to $\text{logit}^{-1}(b_0 + b_1 x)$ when there is variation in ε. Therefore, the variance in ε determined by σ^2 will affect the estimates of b_0 and b_1.

Technically, estimating σ^2 when no phylogenetic signal exists represents a problem of statistical "identifiability" (Judge et al. 1985) because σ^2 and the regression coefficients cannot be estimated simultaneously: Different combinations of parameter values can give an identical fit of the statistical model to the data. Identifiability is a very general phenomenon that plagues statistical analyses when the model is improperly specified. A familiar example is that of collinearity of independent variables in standard linear regression: If two independent variables are perfectly correlated, then it is impossible to estimate regression coefficients for both. This problem technically goes away if the independent variables are even slightly less than perfectly correlated, although practically the problem remains for even moderately highly correlated independent variables (e.g., >0.7) unless sample sizes are large. Our particular identifiability problem is similar in that if any phylogenetic covariances (off-diagonal elements) are present in matrix $\mathbf{C}$, then it is technically possible to estimate σ^2 and the regression coefficients. Nonetheless, in practice, when phylogenetic signal is weak (off-diagonal elements of

matrix $\mathbf{C}$ are small), it will be hard to estimate σ^2 unless sample sizes (number of taxa) are very large. We will discuss statistical problems associated with identifiability when describing our simulation results.

An additional problem arises when comparing values of regression coefficients b_0 and b_1 in Eq. (9.2) for data sets that differ in phylogenetic signal and hence σ^2. Because σ^2 affects the estimates of b_1, apparent differences in regression coefficients among models could be caused by differences in the estimated strength of the relationship between Y and x or by differences in the magnitude of phylogenetic signal. A simple way to at least approximately correct for this is provided by Hadfield (2012), following Diggle et al. (2004): If b_1 is estimated when $\sigma^2 > 0$, then the value that would have been estimated in the absence of variance in ε is approximately $b_1(1 + c^2\sigma^2)^{-0.5}$ where $c = (16/15)(3^{1/3}/\pi)$. Thus, if all estimates are corrected by a factor $(1 + c^2\sigma^2)^{-0.5}$, then the values of the regression coefficients can be compared more directly. Note that although this correction facilitates comparisons among estimates for the PGLMM model (Eq. 9.2), the structure of this model is different from that of the PLogReg model, and therefore, the regression parameters are not expected to have exactly the same values for a given data set, even though both models provide valid estimates of the same relationship between Y and x.

In principle, various approaches can be used for parameter estimation for this PGLMM, although none is easy to implement. For example, the software package ASReml can be configured for PGLMM (Jarrod Hadfield, pers. comm.). Here, we use the approach presented in Ives and Helmus (2011) for solving a more general formulation of PGLMM models designed for data sets containing the presence/absence of species from ecological communities as dependent variables. Parameter estimation involves combining penalized quasi-likelihood (PQL) and restricted maximum likelihood (REML) in a two-step process. This approach gives approximate standard errors for the regression coefficients from which confidence intervals can be calculated and statistical tests can be performed. Bootstrapping can and should also be performed as described above for PLogReg. We implemented PGLMM in MATLAB and provide a version in R in the online practical material (http://www.mpcm-evolution.com); a more general but harder-to-use function is also available in **picante** {R} (Kembel et al. 2010).

9.2.3 Generalized Mixed Model with Bayesian Estimation (MCMCglmm)

A very similar model to PGLMM can be implemented using MCMCglmm {R} (Hadfield 2010) with a Bayesian framework (Chap. 10). The model is

$$\begin{aligned} \Pr(Y = 1) &= p \\ \mathrm{probit}(p) &= b_0 + b_1 x + s + u \\ \mathbf{s} &\sim \mathcal{N}(\mathbf{0}, \sigma_{\mathbf{s}}^2 \mathbf{C}) \\ \mathbf{u} &\sim \mathcal{N}(\mathbf{0}, \sigma_{\mathbf{u}}^2 \mathbf{I}) \end{aligned} \tag{9.3}$$

This is identical to Eq. (9.2), except a probit rather than logit transform is used, and there is an additional random term u. In the general formulation of GLMMs that MCMCglmm is designed to analyze, the random variable u represents residual variation, and s is a "random effect" that captures hypothesized covariances in the data. This model is identical in structure to the GLMMs that can be analyzed using such popular programs as lmer in **lme4** {R}, although most programs (including lmer) for technical reasons restrict the structure of covariance matrix $\mathbf{C}$ to contain blocks of covariance terms rather than covariances specific (and potentially unique) to pairs of species; this means that a phylogenetic covariance matrix of arbitrary form cannot be used. MCMCglmm has no such restriction and is an easy-to-use program for investigating phylogenetic GLMMs.

MCMCglmm is a Bayesian statistical approach, and therefore, statistical results have a different interpretation than those produced by other methods we investigated (Chap. 10). Although long discussions in the literature argue the benefits of Bayesian versus frequentist approaches, most statisticians we know are comfortable using both, adopting Bayesian analyses mainly when the likelihood function does not lend itself to be handled easily using frequentist approaches. The growth in the use of Bayesian statistics has been fueled largely by the growth of inexpensive computing power, which allows the application of the Markov chain Monte Carlo (MCMC) algorithm to complex likelihood functions (Gelman et al. 1995).

MCMC Bayesian statistics, including MCMCglmm, approximate the distribution of a parameter estimate conditional on the observed data and an initial "prior" specification of the parameter distribution before information about the data is used. The mode of the distribution of parameter values from the Markov chain is then the "best" estimate of the parameter, and the spread of the distribution determines the credibility of the parameter estimate. Rather than confidence intervals, MCMCglmm generates credible intervals that give, for example, the range of values that a parameter takes with 95 % probability. Although Bayesian approaches do not give tests of hypotheses in the frequentist sense, the credible intervals give similar information. For example, if the 95 % credible interval of b_1 lies above and does not include zero, then we could say that the value of b_1 is greater than zero with 95 % credibility.

Owing to the structure of MCMCglmm, it is necessary to include the random variable u; we fixed the variance $\sigma_u^2 = 1$, while estimating the variance σ_s^2 as a measure of phylogenetic signal. The inclusion of u makes the model different from the PGLMM model of Eq. (9.2). Nonetheless, this is not a serious difference, because of the identifiability issue discussed for PGLMM (Sect. 9.2.2) when the off-diagonal elements in covariance matrix $\mathbf{C}$ are all zero; the variance in u will

affect the specific values of the estimates of the regression coefficients, but will have little effect on the fit of the model. To make the regression coefficients more comparable between models and data sets, we also correct the regression coefficients of the MCMCglmm model by a factor $(1 + c^2\sigma_s^2 + c^2\sigma_u^2)^{-0.5}$ to "take away" the effect of variances on the estimates of the regression coefficients (Hadfield 2012); this should make values more comparable to those from PGLMM.

To run MCMCglmm, it is necessary to make decisions about the structure of the model, the prior parameter distributions, and the run characteristics. In our simulations, we found that using a probit transform (Eq. 9.3) performed better than a logit transforms as used in PGLMM (Eq. 9.2). We followed the recommendations of Hadfield (2012, Sect. 8.0.8) to use slice sampling and expanded priors for variance parameters σ_s^2 and σ_u^2, choosing X^2 priors as suggested by Villemereuil et al. (2013). The MCMCglmm defaults were used for the length of the burn-in (3,000) and sampled (10,000) chain lengths, with 1/10 samples taken for the posterior parameter distributions. Diagnostics (Hadfield 2012) showed adequate chain mixing for these settings. Discussion of these assumptions is provided in the online practical material (http://www.mpcm-evolution.com).

9.2.4 Regression with a Model-Dictated Branch-Length Transformation (RegOU)

For continuously valued dependent variables, the current state of the art in phylogenetic analyses is to perform regression while simultaneously estimating the strength of phylogenetic signal using a branch-length transform (Chap. 15). This is possible in a variety of packages; for example, **ape** {R} (Paradis et al. 2004) provides phylogenetic correlation structures corresponding to different models of evolution that can be used in regression through the gls function of **nlme** {R}. As noted elsewhere, so-called PGLS methods are equivalent to phylogenetically independent contrasts in their simplest form and do not inherently include estimation of branch-length transforms (Lavin et al. 2008). Here, we use the module RegOU in the Regressionv2.m (Lavin et al. 2008) MATLAB program to perform equivalent analyses under the assumption that evolution follows an Ornstein–Uhlenbeck process, intended to mimic stabilizing selection (Felsenstein 1988; Garland et al. 1993; Chap. 15). The strength of stabilizing selection given by the parameter d determines the phylogenetic signal estimated to exist in residuals. In the absence of stabilizing selection ($d = 1$), RegOU gives BM evolution, while stronger stabilizing selection (smaller d) erases evolutionary memory and reduces phylogenetic signal. When $d = 0$, phylogenetic signal disappears, and the estimated model parameters are then identical to those obtained by an ordinary least-squares analysis, which in effect assumes a star phylogeny with no hierarchical structure. An advantage of integrating a branch-length transform into the

regression analysis is that an a priori decision about the existence of phylogenetic signal is unnecessary; instead, it is estimated along with the regression coefficients—the data are allowed to tell their own story (Lavin et al. 2008).

We used RegOU even though we know the binary data Y violate the RegOU assumption that the dependent variable takes continuous values. Although one could object to applying a statistical model that is known to be wrong, all statistical models are likely to be wrong in some way to some degree, and the assumption of RegOU that Y be continuous might not make RegOU useless. By the central limit theorem, even a process as discrete as flipping a coin will converge to a Gaussian distribution if repeated enough times. With a sufficient number of taxa, RegOU might be adequate to identify, for example, the existence of an effect of independent variable x on Y. Because RegOU does not produce a model with discrete outcomes, however, the values of the regression coefficients are difficult to interpret. Furthermore, it is not possible to use a fitted model to simulate data, and therefore, it is not possible to perform bootstrapping. Nevertheless, RegOU might perform reasonably well for hypothesis testing (i.e., yield reasonable Type I errors and power).

9.2.5 Ancestral Character Estimation (ACE)

Pagel (1994) presented a rapid method for computing the exact likelihood for a model of two binary traits evolving along a phylogenetic tree in a correlated fashion. Using the same approach (Schluter et al. 1997) but applied to a single trait gives a method for estimating phylogenetic signal for a single binary trait. For a single trait, the assumed evolutionary process is essentially identical to the first part of the PLogReg model that generates phylogenetic correlations (Sect. 9.2.1). The trait Y has an instantaneous probability of changing from 0 to 1 and from 1 to 0, and the longer the branch lengths between nodes, the greater the probability of switches (and back-switches). The program ACE in **ape** {R} (Paradis et al. 2004) computes the ML of this process while fitting the transition rate parameters.

Because ACE does not lead to a model that can be used for regression with an independent variable x, we only use it to detect phylogenetic signal. As with PLogReg, higher transition rates reduce phylogenetic signal and therefore could be used to assess phylogenetic signal. Nonetheless, this does not lead to a clear test of whether phylogenetic signal is present, because the threshold for the transition rates above which we should declare no signal is unclear. Therefore, we used the following procedure. We used Grafen's (1989) rho branch-length transform (that is easily implemented using the compute.brlen function in **ape** {R}) to modify a given topology and hence produce trees representing different degrees of expected phylogenetic signal. As rho approaches zero, the tree becomes a star, with no phylogenetic covariances between taxa; in compute.brlen, rho can never equal zero, so we used a minimum values of rho $= 10^{-5}$ instead of zero. (Note that this restriction on the lower limit of rho is particular to compute.brlen; other programs

allow rho to become zero [e.g., DOS PDTREE and the PDAP module of Mesquite].) Larger values of rho correspond to greater shared branch lengths and hence greater phylogenetic structure. We used ACE to calculate the ML of the model over the possible transition rates for a given value of rho and then selected that value of rho giving the greatest likelihood. Thus, a test for the existence of phylogenetic signal is whether the rho giving the greatest likelihood is statistically greater than zero. ACE can analyze several models of evolution, such as allowing a different rate for the 0 to 1 transition than for the 1 to 0 transition. However, in preliminary simulation experiments, we found that the simplest model assuming transitions are symmetrical gave the greatest power to detect phylogenetic signal.

9.2.6 Standard Logistic Regression with a Firth Correction (Logistf)

To compare with a non-phylogenetic analysis, we used standard logistic regression with a Firth correction to reduce bias in the estimates (Ives and Garland 2010) as implemented by **logistf** {R} (Heinze et al. 2013). We only used this model in simulations including an independent variable x.

9.3 Method Comparison Using a Simulation Model

We performed simulations both without and with an independent variable x. In the absence of x, the phylogenetic models become tests of phylogenetic signal. To generate simulation data for comparisons among models, we chose a model similar to that underlying PLogReg, although differing slightly in the manner of generating phylogenetic signal. We first simulated switches in the value of Y along a phylogenetic tree at a rate given by $a = 0$, which gives phylogenetic signal comparable in magnitude to BM evolution of continuous traits (Ives and Garland 2010). However, rather than vary the value of a to change the strength of phylogenetic signal, instead, we performed an OU transform on the phylogenetic tree (following the parameterization of Lavin et al. 2008), with $d = 0$ giving no phylogenetic signal (a star phylogeny) and increasing values of d giving greater phylogenetic signal (see Sect. 9.2.4 above and Chap. 15 for a description of the OU branch-length transform). For the case with an independent variable x, again as in PLogReg, we assumed that a second phase of evolution occurred in which values of Y (at the tips of the tree) tended to switch toward 0 or 1 depending on the value of x and the regression coefficient b_1, and independently from the phylogeny.

To simulate values of x, we assumed values were chosen either with or without phylogenetic signal. To model phylogenetic signal in x, we assumed BM evolution

on the phylogenetic tree leading to a multivariate Gaussian distribution with mean 0, variance 1, and covariances proportional to the shared branch length for each tip species. To model x in the absence of phylogenetic signal, we chose values of x independently from a Gaussian distribution with mean 0 and variance 1. We considered only "symmetrical" ("balanced") phylogenetic trees; preliminary simulations showed little difference in results between symmetrical and highly asymmetrical "ladder" trees (see also Blomberg et al. 2003). Finally, when simulated data sets had $\geq 7/8$ of the values of Y all zeros or all ones, we discarded them because they will contain little information, and a prudent researcher would not analyze them in the first place (cf. Diaz-Uriarte and Garland 1996, p. 45).

9.3.1 Phylogenetic Signal (Regression Without Independent Variables)

We first investigated the ability of five methods—PLogReg, PGLMM, MCMCglmm, RegOU, and ACE—to detect phylogenetic signal. We simulated 1,000 data sets for 64 species on a symmetrical phylogenetic tree subjected to an OU branch-length transform. We used d values in increments of 0.1 in the range of 0 to 0.8; although d can exceed unity, values >0.8 were deemed unnecessary, given the results obtained. For each of the five methods, we plotted the mean estimates of the respective phylogenetic signal parameters and their 66 and 90 % inclusion intervals (Fig. 9.1). The 66 % inclusion interval is comparable to ± 1 standard deviation. The 90 % inclusion interval gives an indication of the power of each method to reject the null hypothesis of no phylogenetic signal; because this hypothesis is one-sided, the lower boundary of the 90 % inclusion interval corresponds to 5 % of the simulated data having estimates of no phylogenetic signal. It is important to note that these inclusion intervals are not confidence intervals, because the simulations were produced under a model that differed from all of the statistical models. Nonetheless, the inclusion intervals give an indication of the probability that the point estimate of phylogenetic signal from each model is greater than zero. We did not perform simulations that involved estimating confidence intervals due to the computational intensity required; for example, if confidence intervals were computed for 1,000 bootstrap data sets for each of the 1,000 simulations, 1,000,000 estimations would be needed at each level of phylogenetic signal.

PLogReg was the best method for detecting signal, followed by PGLMM. Given the similarity between the PGLMM and MCMCglmm models (Eqs. 9.2 and 9.3), we were surprised that MCMCglmm did not perform better; even for the simulation $d = 0.8$, more than 12 % of the simulations resulted in MCMCglmm estimates of $\sigma_s^2 = 0$ indicating no phylogenetic signal (Fig. 9.1c). We suspect that this involves the identifiability problem that affects both PGLMM and MCMCglmm. When no phylogenetic signal exists (zero covariances in Y among taxa), the parameters b_0 and σ^2 (σ_s^2 for MCMCglmm) are confounded by the identifiability problem; no
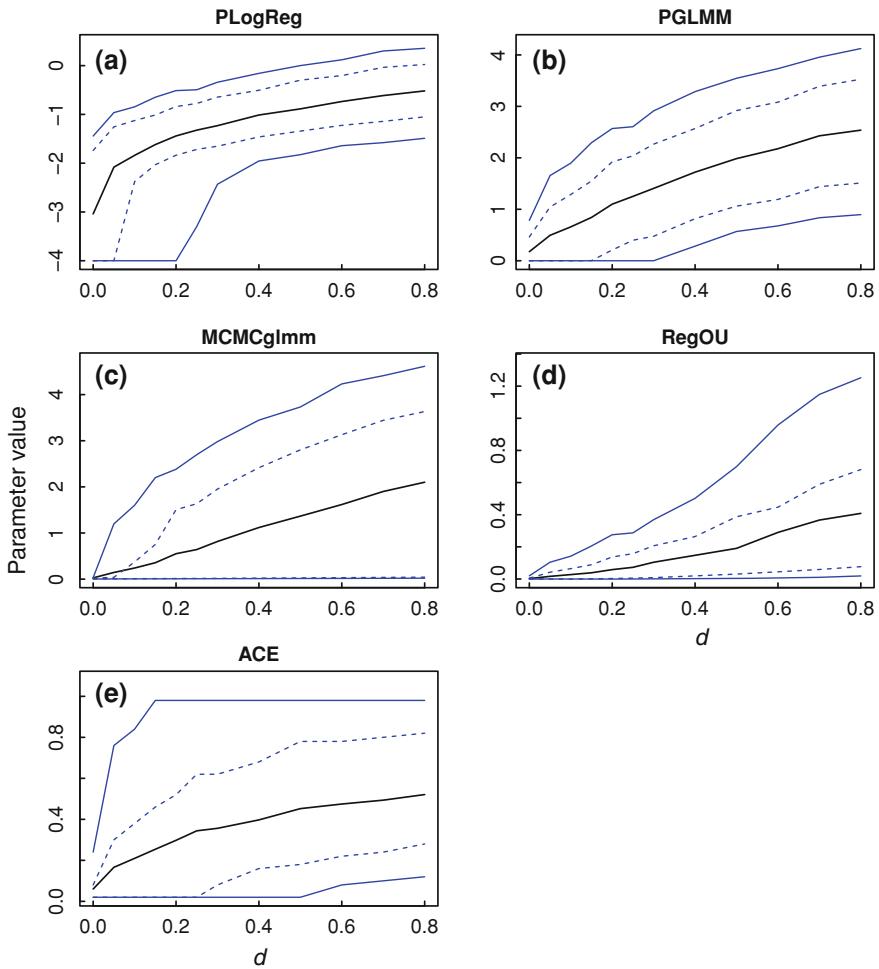
Fig. 9.1 Estimates of phylogenetic signal from **a** PLogReg, **b** PGLMM, **c** MCMCglmm, **d** RegOU, and **e** ACE. In all cases, the horizontal axis gives d of the OU branch-length transform used to general different relative branch lengths for phylogenetic trees up which the binary dependent variable evolved. The vertical axis gives the phylogenetic signal parameter for each of the five methods. In each panel, the central line is the mean parameter values, and *dashed* and *solid lines* give the 66 and 90 % inclusion intervals from 1,000 simulations at each value of d with $a = 0$. Phylogenetic trees were all assumed to be symmetrical with 64 taxa. The lower boundary of the 90 % inclusion interval corresponds to 5 % of the simulated data having estimates of no phylogenetic signal, which gives an indication of the probability that the point estimate of phylogenetic signal from each model is greater than zero. These simulations show the relative performance of the methods at identifying phylogenetic signal, with PLogReg outperforming the other methods, and able to detect signal when the value of d equals or exceeds 0.2. Note that d can exceed unity, but simulations with d set at values > 0.8 were deemed unnecessary

single pair of values gives the best model fit. The estimation approach of PGLMM uses an iterative process alternating conditional likelihoods for b_0 given σ^2 and for σ^2 given b_0, and we suspect that this acts to separate (i.e., decrease the correlation) of the estimates of these parameters. By decreasing this correlation, PGLMM will reduce the variability in the estimate of σ^2. Consistent with this explanation, for simulations with $d = 0.5$, the correlation between the estimates of $|b_0|$ and σ for PGLMM was -0.26, while for MCMCglmm the correlation between $|b_0|$ and σ_s^2 was -0.35. At $d = 0.5$, zero is well outside the 90 % inclusion interval for PGLMM while still within the inclusion interval for MCMCglmm.

RegOU did not perform well. This is presumably because it did a poor job capturing the phylogenetic covariances among values of Y. By ignoring the binary nature of the data, RegOU does not incorporate the constraints on the variances and covariances that are imposed by the nature of binary data. Specifically, the covariances are bounded by the variances in Y, and the variances in turn are set by $p(1 - p)$, which has a maximum value of $p = 0.5$. Therefore, if by chance there is a relatively large number of ones (or zeros) in a data set, then the estimate of p will be greater (or less) than 0.5 and the covariances in the data will necessarily be reduced. RegOU does not account for this, and we suspect that this is the reason for its poor power. In contrast, PLogReg explicitly accounts for the dependency of the variances and covariances on p, in effect increasing the weight of the covariances in the data to compensate for changes in p.

The performance of ACE was third after PGLMM. We expected ACE to perform better. The performance of ACE could be limited by its use of ML, in contrast to penalized likelihood used by PLogReg and REML used by PGLMM, although other explanations are possible, including the mismatch between the model underlying ACE and the simulation model that we used.

9.3.2 Regression

We investigated the performance of five methods—PLogReg, PGLMM, MCMCglmm, RegOU, and Logistf—to estimate b_1 giving the relationship between Y and the independent variable x. We simulated 1,000 data sets with $b_0 = 0$ and $b_1 = 1$ for 64 species on a symmetrical phylogenetic tree subjected to an OU branch-length transform with values of d between 0 and 2 in increments of 0.25. The values of x were assumed to have evolved under a BM model up the phylogenetic tree with no branch-length transform (i.e., $d = 1$). For each data set, we then estimated b_1 using all five methods and plotted the means from the 1,000 simulations, as well as the 66 and 90 % inclusion intervals (Fig. 9.2).

Because of differences between the simulation and statistical models, we did not expect the estimates of b_1 to be exactly 1. Nonetheless, changes in the estimates of b_1 with d used to simulate the data indicate that phylogenetic signal introduces bias in the estimates. All methods except RegOU showed upward bias
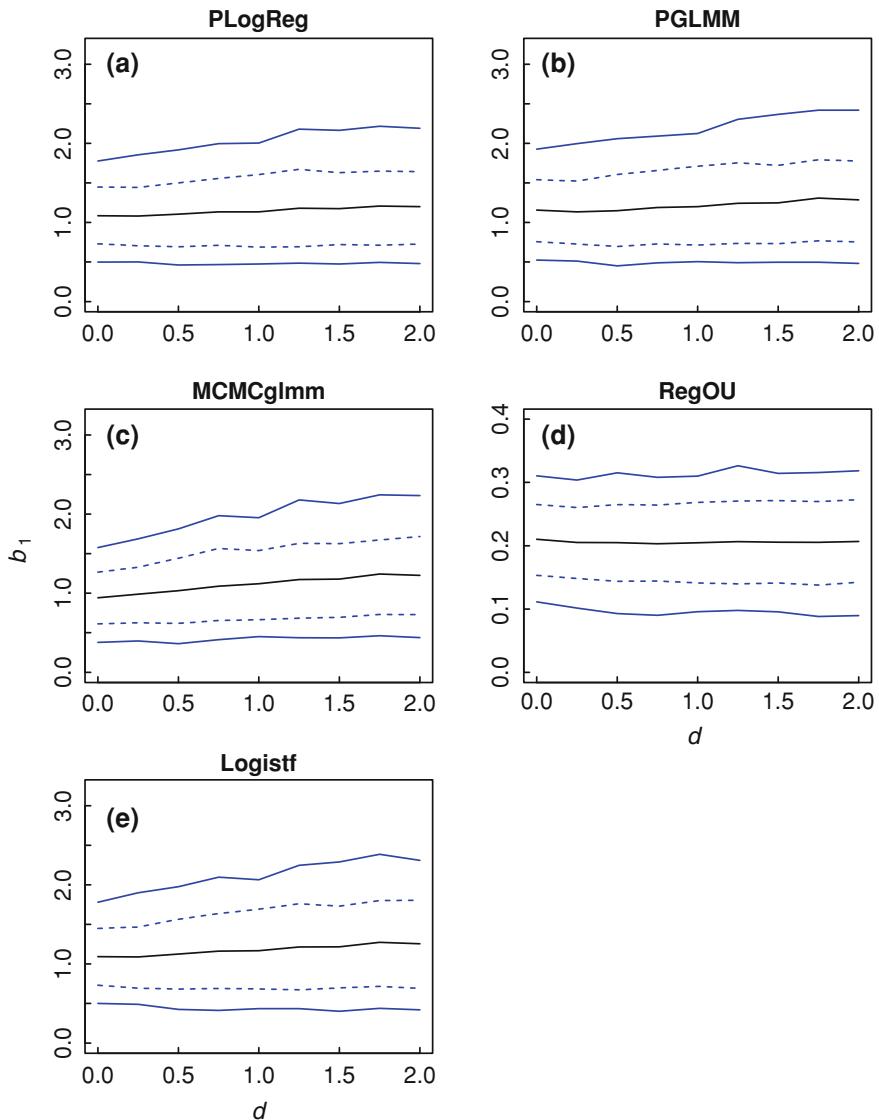
Fig. 9.2 Estimates of regression parameter b_1 from **a** PLogReg, **b** PGLMM, **c** MCMCglmm, **d** RegOU, and **e** Logistf. In all cases, the horizontal axis gives d of the OU branch-length transform used to generate different relative branch lengths for phylogenetic trees up which the binary dependent variable evolved, and in the simulations $b_1 = 1$. In each panel, the central line is the mean parameter values, and *dashed* and *solid lines* give the 66 and 90 % inclusion intervals from 1,000 simulations at each value of d with $a = 0$. Values for PGLMM and MCMCglmm were corrected by factors $(1 + c^2\sigma^2)^{-0.5}$ and $(1 + c^2\sigma_s^2 + c^2\sigma_u^2)^{-0.5}$, respectively (see text). For simulations, the independent variable x was assumed to evolve as a Brownian motion process along the specified tree. Phylogenetic trees were all assumed to be symmetrical with 64 taxa. These results show both increased bias (mean parameter values) and decreased precision (width of inclusion intervals) with increasing phylogenetic signal d

with increasing phylogenetic signal, with PLogReg showing the least (12 % when $d = 2$ vs. $d = 0$) and MCMCglmm showing the most (32 % when $d = 2$ vs. $d = 0$). Simultaneously, increasing d made the estimates of b_1 more variable for all methods, as shown by the broadening of the inclusion intervals. This loss of precision was least for RegOU (15 %) followed by PLogReg (34 %) and was greatest for Logistf (48 %). Thus, failing to account for phylogenetic signal in Logistf led to the greatest loss in precision as phylogenetic signal increased.

One of the greatest statistical concerns driving phylogenetic comparative methods is Type I errors, rejecting the null hypothesis when it is true (e.g., see Grafen 1989; Martins and Garland 1991; Garland et al. 1993; Diaz-Uriarte and Garland 1996). To investigate Type I errors, we simulated data under the null hypothesis of $b_1 = 0$ and scored for each statistical method the proportion of simulations for which the null hypothesis was rejected at the $\alpha = 0.05$ significance level (Fig. 9.3). The data were simulated up balanced, ultrametric phylogenetic trees for $N = 16, 32, 64, 128$, and 256 terminal taxa assuming substantial phylogenetic signal ($d = 1$). Values of the independent variable x were simulated either with phylogenetic signal given by BM evolution (Fig. 9.3a) or without phylogenetic signal (Fig. 9.3b). The statistical tests for $H_0:b_1 = 0$ were performed using asymptotic approximations with PLogReg, PGLMM, RegOU, and Logistf. For MCMCglmm, to give comparable information to hypothesis testing, we used the 95 % credible interval for b_1, scoring the proportion of simulations for which zero fell outside this interval.

Ideally, all methods would reject $H_0:b_1 = 0$ in 5 % of the simulated data sets when the value of b_1 in the simulations is in fact zero. Because we only performed 1,000 simulations, there will be some variability around this 5 % expectation. Specifically, for 1,000 simulations, the 95 % confidence interval for expected number of simulated data sets rejecting $H_0:b_1 = 0$ when the rejection rate is correct can be calculated to be (3.6, 6.4 %); consistent departures from this range imply incorrect Type I error rates. When phylogenetic signal exists in the independent variable x, Logistf rejects $H_0:b_1 = 0$ much more frequently, indicating an inflated Type I error rate (Fig. 9.3a). In the absence of phylogenetic signal in x, however, Logistf does not perform badly (Fig. 9.3b). This type of behavior—sensitivity of a statistical regression method to the distribution of the independent variable—is also found in phylogenetic regression with continuous dependent variables (Revell 2010). In standard regression, when phylogenetic signal is absent in the residual variation, the distribution of the independent variables does not affect the estimates and statistical tests of the corresponding regression coefficients. However, with phylogenetic signal in the residuals, the presence or absence of phylogenetic signal in the independent variables does matter. The results for Logistf show that when an independent variable is correlated with the residuals, the data contain less information about the parameters. Logistf ignores this loss of information, thereby giving incorrectly strong statistical tests leading to badly inflated Type I error rates.
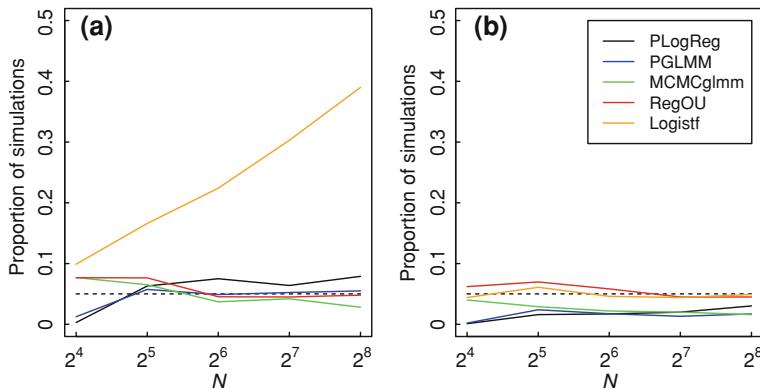
Fig. 9.3 Type I error rates (proportions of simulations in which H_0: $b_1 = 0$ was rejected when it was true) at $\alpha = 0.05$ as a function of the number of taxa in the sample (16, 32, 64, 128, and 256) on a symmetrical phylogenetic tree for PLogReg (*black*), PGLMM using approximate standard error (*blue*), MCMCglmm (*green*), RegOU (*red*), and Logistf (*orange*). *Dotted black line* indicates nominal Type I error rate of 5 %. In **a** the independent variable x is assumed to show Brownian motion phylogenetic signal, and in **b** x is assumed to have no phylogenetic signal (evolution up a star phylogeny). In both **a** and **b** there is phylogenetic signal in the residual variation given by $d = 1$ and $a = 0$. For all methods, the lines give the proportion of 1,000 simulations in which the null hypothesis $b_1 = 0$ is rejected at the $\alpha = 0.05$ level based on the approximate asymptotic distribution of b_1, except for MCMCglmm which gives the proportion of simulations in which the 95 % credible interval excludes $b_1 = 0$. Simulated data sets with Y taking few values of 0 or 1 ($\leq 1/8$ values) were excluded, because these data sets will give little information for statistical fitting and a practitioner probably should not try to analyze them statistically (e.g., see Diaz-Uriarte and Garland 1996, p. 45). Results in this figure show that all methods other than PGLMM (*blue line*) show inflated Type I error rates for some values of N when there is phylogenetic signal in x (panel **a**)

The phylogenetic methods performed much better than Logistf when phylogenetic signal is present in x (Fig. 9.3a). For small sample sizes ($N = 16$), both PLogReg and PGLMM underestimated the nominal number of simulations with "significant" values of b_1, although they performed well for $N \geq 32$; this underestimate of significance (lower-than-appropriate Type I error rates) when $N = 16$ is conservative, in the sense that it means that false positives are less likely. MCMCglmm had inflated Type I error rates for $N = 16$, which is an issue of concern, although MCMCglmm performed similarly to PLogReg and PGLMM for $N \geq 32$. Overall, RegOU performed well although, like MCMCglmm, it showed inflated Type I error for $N = 16$. Without phylogenetic signal in x (Fig. 9.3b), PLogReg, PGLMM, and MCMCglmm all showed lower-than-appropriate Type I error rates, especially PLogReg and PGLMM when $N = 16$. Perhaps most surprisingly, the two best performers were RegOU and Logistf that ignore, respectively, the binary nature of the data and phylogenetic signal. These results illustrate that, in statistics, it is sometimes possible for a method to be right even when it does not incorporate all of the characteristics of the data.

To investigate statistical power, we generated curves of the probability of rejecting the null hypothesis $H_0:b_1 = 0$ with values of b_1 ranging from 0 to 1 (Fig. 9.4). The data were simulated for 64 species assuming phylogenetic signal in Y ($d = 1$). Values of the independent variable x were simulated either with phylogenetic signal given by BM evolution (Fig. 9.4a) or without phylogenetic signal (Fig. 9.4b). For PGLMM, in addition to performing the analyses using the asymptotic approximation for standard errors of b_1, we also performed a bootstrap test of $H_0:b_1 = 0$ using the 95 % bootstrap confidence intervals of b_1. To make this feasible, we used only 200 bootstrapped data sets for each of 200 simulated initial data sets, which still required 40,000 estimations for each of the 11 values of d. We did not perform a similar test of the bootstrap for PLogReg, because the computer time to run the 440,000 estimations was prohibitive. For single, real data sets, however, bootstraps can be run for PLogReg.

RegOU performed well, with a correct Type I error rate (at $d = 0$) and good power, especially when phylogenetic signal is absent in x (Fig. 9.4b). With phylogenetic signal in x (Fig. 9.4a), MCMCglmm had greatest power, at least for this sample size of $N = 64$ (but see Fig. 9.3a). In part, this might be due to the greater upward bias shown by MCMCglmm (Fig. 9.2c), which makes it more likely for the 2.5 % tail of the distribution of the estimate of b_1 to lie above zero. Using the asymptotic approximations for the standard error of b_1, both PLogReg and PGLMM had slightly lower power than MCMCglmm and comparable power to RegOU. The bootstrap results for PGLMM had the correct Type I error rates at $d = 0$; correct Type I error rates are guaranteed with bootstrapping, provided the data are well fit by the PGLMM model, because bootstrapping is based on simulating data and scoring those for which $H_0:b_1 = 0$ is rejected. Despite having the correct Type I error rates when $d = 0$, the PGLMM bootstrap had lower power than the other methods. Despite this low power, because the bootstrapping guarantees lack of Type I errors, it provides the most "secure" results.

Overall, the power curves show that all methods have roughly similar power. However, given the inflated Type I error rates shown by all methods except for PGLMM for some sample sizes (Fig. 9.3a), all phylogenetic methods run the risk of giving false positive results. Generally, Type I errors are of greater statistical concern than Type II errors (accepting the null hypothesis when it is false), and the Type I errors shown by the simulations should lead to caution when interpreting results. Indeed, many statisticians would not move on to consider statistical power unless Type I error rates were assured at a nominal level, such as $\alpha = 0.05$ (e.g., Martins and Garland 1991).

9.4 Method Comparison with Real Data

To compare methods using real data, we used an example provided by Brashares et al. (2000) for 75 species of African antelope; this data set was also analyzed using PLogReg in Ives and Garland (2010). We tested the hypothesis proposed by
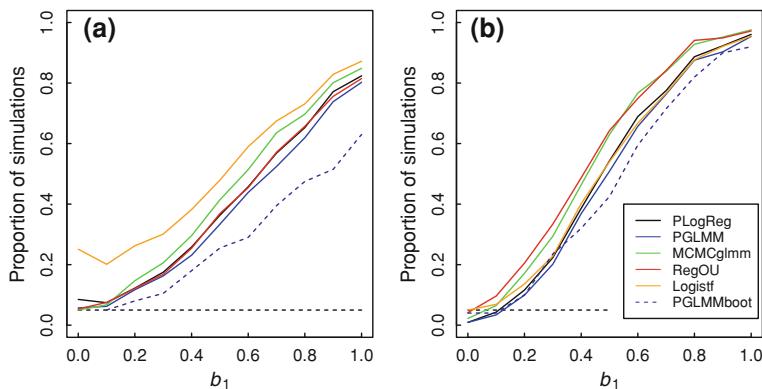
Fig. 9.4 Power curves for estimates of the regression parameter b_1 from PLogReg (*black*), PGLMM using approximate standard error (*blue*), PGLMM using bootstrapped confidence intervals (*dashed dark blue*), MCMCglmm (*green*), RegOU (*red*), and Logistf (*orange*). The power curves give the proportion of simulations in which the null hypothesis H_0: $b_1 = 0$ was rejected at the $\alpha = 0.05$ level for 64 taxa evolving on a symmetrical phylogenetic tree. In **a** the independent variable x is assumed to show Brownian motion phylogenetic signal, and in **b** x is assumed to have no phylogenetic signal (evolution up a star phylogeny). In both **a** and **b**, there is phylogenetic signal in the residual variation given by $d = 1$ and $a = 0$. For all methods, the lines give the proportion of 1,000 simulations in which the null hypothesis H_0: $b_1 = 0$ was rejected at the $\alpha = 0.05$ level, except for MCMCglmm which gives the proportion of simulations in which the 95 % credible interval excludes $b_1 = 0$, and the bootstrapped results for PGLMM that are based on 200 simulations. These results show that all methods have similar power, with the bootstrapped power curve for PGLMM having the lowest power

Jarman (1974) that species living in larger groups are more likely to flee or fight predators ($Y = 1$), whereas solitary or pair-living species are more likely to hide ($Y = 0$). Group size ranges between 1 and 70, and we treated $\log_{10}$-transformed group size as a continuous variable. Because body size is likely also to affect antipredator behavior, with larger-bodied species more likely to flee/fight than hide, we followed Brashares et al. (2000) and also included $\log_{10}$ body mass as a second independent, continuously valued variable. To stabilize the statical analyses, both independent variables were standardized to have mean equal to zero and standard deviation equal to one; this also makes the regression coefficients represent effect sizes of the independent variables whose magnitudes reflect the size of effect of the variable (as is traditionally done in path analysis, Chap. 8). Group size had phylogenetic signal (RegOU: $d = 0.84$, conf. interval = (0.52, 1.07); Lavin et al. 2008), as did body size ($d = 0.99$, conf. interval = (0.77, 1.20)), suggesting that phylogenetic signal in Y will present statistical challenges. Here, we summarize the results of the analyses, and in the online practical material (http://www.mpcm-evolution.com), we present this example as a tutorial.
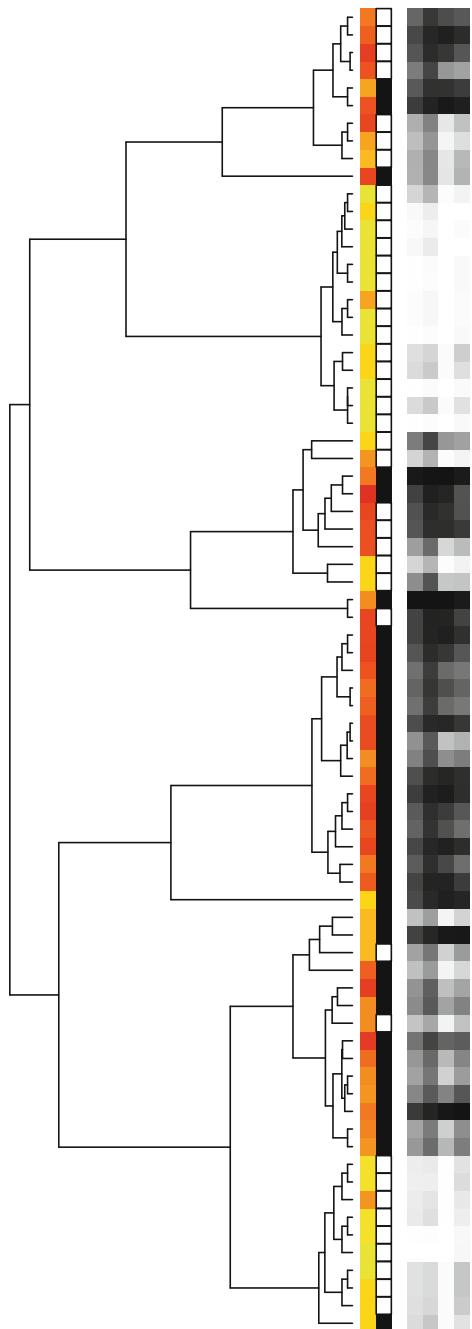
All methods revealed a strong positive effect of group size on the response of antelope to predators (Table 9.1, Fig. 9.5); antelopes with larger group sizes were more likely to flee/fight than to hide. Because the methods incorporate the regression

Table 9.1 Comparison among five methods of estimating regression coefficients for the effects of $\log_{10}$ group size and $\log_{10}$ body mass on the antipredator behavior (0 = hide, 1 = flee or fight) of 75 antelope species

Parameter[a]	Estimate	Approx. SE[b]	95 % confidence/ credible interval[c]	p-value	Bootstrap mean[d]	95 % bootstrap confidence interval
PLogReg						
a	0.50				−0.49	(−4, 4)
b_0	−0.82	0.87	(−2.54, 0.90)	0.34	−0.86	(−3.67, 1.53)
b_1 (body mass)	0.096	0.45	(−0.80, 0.99)	0.84	0.13	(−1.08, 1.34)
b_2 (group size)	1.36	0.49	(0.39, 2.33)	0.007	1.71	(0.47, 3.66)
PGLMM						
σ^2	7.16				2.86	(0.00, 8.0)
b_0	−0.70	0.79	(−2.27, 0.87)	0.38	−1.13	(−3.27, 0.35)
b_1 (body mass)	0.22	0.60	(−0.98, 1.42)	0.71	0.51	(−0.98, 2.20)
b_2 (group size)	1.45	0.66	(0.13, 2.77)	0.031	1.66	(0.56, 3.24)
MCMCglmm						
σ_s^2	4.28		(1.50, 13.9)		0.83	(0.01, 3.28)
σ_u^2	1				1	
b_0	−5.23		(−13.2, −1.12)		−3.92	(−10.1, −0.34)
b_1 (body mass)	0.54		(−2.56, 4.03)		0.49	(−1.81, 3.23)
b_2 (group size)	4.20		(1.50, 13.9)		3.07	(0.68, 6.55)
RegOU						
d	0.62					
b_0	0.39	0.20	(−0.008 0.79)	0.47		
b_1 (body mass)	0.16	0.070	(0.021, 0.30)	0.026		
b_2 (group size)	0.16	0.056	(0.049, 0.27)	0.007		
Logistf						
b_0	−1.57	0.96	(−3.47, 0.33)	0.09	−1.59	(−3.66, 0.18)
b_1 (body mass)	−1.17	0.87	(−2.89, 0.55)	0.16	−1.16	(−3.01, 0.39)
b_2 (group size)	4.57	1.25	(2.10, 7.05)	<0.0001	4.56	(2.27, 7.44)

[a] All independent variables were standardized to have mean 0 and variance 1 prior to analysis
[b] Standard errors of the estimates and confidence intervals were obtained using the asymptotic approximations
[c] Approximate confidence intervals for the frequentist methods were computed using asymptotic approximations. The credible intervals for MCMCglmm were computed in the fitting process
[d] Parametric bootstrapping was performed by simulating 1,000 data sets to obtain means and confidence intervals

Fig. 9.5 Analyses of antipredator behavior ($0 =$ hide, $1 =$ flee or fight) of 75 antelope species as it depends on $\log_{10}$ group size (see results shown in Table 9.1). The *first column* to the *right* of the phylogenetic tree gives $\log_{10}$ group size, a continuously valued independent variable, colored from small (*yellow*) to large (*red*). The second column gives the trait value Y (hide $=$ white, flee/fight $=$ black). The remaining 4 columns give the fit of the models—PLogReg, PGLMM, MCMCglmm, and Logistf—as the probability of fleeing/fighting predicted by the models (Table 9.1) scaled from fleeing/fighting with probability zero (*white*) to one (*black*). These fits also incorporate the information from the second independent variable, $\log_{10}$ body size. All methods revealed the effect of group size on the response variable (Table 9.1), and their similar predictions are shown by similar patterns in the last 4 columns

coefficients differently, their values cannot be compared directly. Nonetheless, all methods showed significant departure of b_2 from zero. Although this main conclusion would be reached regardless of the method used, the results nonetheless show properties of the methods that we will discuss for each method in turn.

For PLogReg, bootstrapping shows that the value of b_2 is an overestimate, because the bootstrap mean is 1.71 compared to the value of 1.36 from the data that were used to parameterize the simulation bootstraps. Furthermore, the bootstrap confidence intervals are wider than those obtained from the approximate standard errors (approximate (0.39, 2.33); bootstrapped (0.47, 3.66)). Simultaneously, the bootstrap confidence interval for a ranging from the minimum to maximum values (-4 to 4) indicates that PLogReg is unable to determine whether there is phylogenetic signal in the residuals.

Like PLogReg, the bootstrap of PGLMM shows upward bias in the estimate of b_2. The bootstrap confidence intervals show no statistically significant phylogenetic signal in the residuals (lower bound for σ^2 is zero). Nonetheless, due to the upward bias in the estimates, the lower bound of the confidence interval could also be upward biased. To test for this, we also performed a bootstrap under the null hypothesis $H_0:b_2 = 0$ by fitting the model without group size x_2, simulating data sets from the fitted model, and then fitting the full model including group size to the simulated data sets. Only 1/1000 simulations had an estimated value of $b_2 > 1.45$ (the value observed in the real data set), implying strong rejection of the null hypothesis $H_0:b_2 = 0$.

Despite using essentially the same statistical models, the parameter estimates from MCMCglmm were substantially different from PGLMM. We produced "bootstrap confidence intervals" for MCMCglmm by simulating 1,000 data sets from the fitted MCMCglmm model and reporting the range covered by 95 % of the estimated parameters. This procedure showed downward bias in the estimate of b_2 although little bias in b_1.

RegOU showed not only a statistically significant effect of group size but also an effect of body size. Based on the simulations that show RegOU gives correct Type I errors and has good power, we are inclined to trust these results with respect to hypothesis testing. Nonetheless, it is not possible to interpret the values of the parameter estimates in a meaningful way; they cannot be converted into the probability that predators flee or fight. This also means that bootstrapping is impossible, leading us to rely on the simulations (Figs. 9.2, 9.3 and 9.4) to give us faith in the results.

Logistf gave, on the face of it, very good statistical properties. The effect of group size was highly significant, and the bootstrapping showed that there is little bias and that the approximate confidence intervals are accurate. However, because both group size and body size show phylogenetic signal, we know that Logistf is likely to suffer from severely inflated Type I errors (Fig. 9.3). Therefore, we cannot trust the statistical results. This shows that even "good" superficial statistical behavior can be underlain by serious statistical mistakes.

Assessing the methods by their fits to the data, all models gave similar predictions for the probability that antelope trait flee/fight (Fig. 9.5). There are clear

cases in which all models predict $Y = 1$ with high probability, yet in fact $Y = 0$ (compare column 2 with columns 3–6). These mis-predictions also appear to have a phylogenetic pattern, with closely related species showing $Y = 0$ despite predictions otherwise. Nonetheless, none of the methods identified statistically significant phylogenetic signal in the residual variation (Table 9.1).

9.5 Discussion

Of the several methods now available for analyzing phylogenetic binary data, our comparisons gave no single "winner" that performed best under all situations and for all questions asked about the data. If we had considered a greater range of phylogenetic trees, sample sizes, and data characteristics (e.g., including multiple regressions), then a "winner" might have been even less apparent. This is not surprising, given the complexity of these methods and of the phylogenetically evolved data they are trying to analyze. Below, we try to give simple guidelines for different situations that might commonly arise, treating first the detection of phylogenetic signal and then regression with independent variables. We emphasize, however, that multiple methods should be tried for any data set; data are generally far more time-consuming to generate than are statistical models to run, so once data are in hand, it makes sense to run multiple tests. If they all give the same results, this is reassuring. If they do not, then our analyses have hopefully pointed to reasons for the differences and helped identify which methods to trust. Nonetheless, we stress that our simulations are far from exhaustive, and it is best to tailor simulations to particular data sets. Moreover, we hope and expect new methods will be developed which can be applied to old data, as raw data are now becoming routinely deposited through online supplementary material and communal repositories.

9.5.1 Phylogenetic Signal (Regression Without Independent Variables)

At least in our simulations, PLogReg was the most powerful method for detecting phylogenetic signal in the absence of independent variables. It was the most likely to show the existence of phylogenetic signal when it was in fact present (Fig. 9.1). On the downside, PLogReg is also the most computer intensive of the methods considered here. For large numbers of taxa (e.g., $N = 256$), bootstrapping can take a day on a typical personal computer. Speed might be considerably helped, however, by the methods of Ho and Ane implemented in **phylolm** {R} (Ho and Ane 2014).

In some situations, other methods are preferable. Following PLogReg, PGLMM performs well and typically runs 100–1,000-fold faster. Therefore, for large data sets or for extensive bootstrapping, PGLMM might be preferred. MCMCglmm, as

a Bayesian method, gives information about the distributions and correlations among parameter estimates, and MCMCglmm as a package gives nice tools that make it possible to visualize and extract this information. This information is helpful for diagnosing problems with model estimation; it can be used in a way similar to bootstrapping to approximate the joint distribution of parameter estimates. An advantage of MCMCglmm is that rather than iterating the estimation procedure many times as required for bootstrapping, MCMCglmm provides diagnostic information in a single estimation, making it functionally very fast.

Unlike the other methods, ACE provides estimates of ancestral states and uncertainty in these estimates. This can be useful to illustrate possible evolutionary sequences of events leading to the current distribution of traits among taxa. Personally, we are cautious about trying to infer much about evolutionary history from information on only extant species (Garland et al. 1999; Bonine et al. 2005; Chap. 22), yet mapping-inferred ancestral trait changes onto phylogenies can be useful in conjunction with historical information about possible drivers of evolution, such as environmental changes that cause alterations in the selective regime (Hansen and Orzack 2005).

9.5.2 Regression

When an analysis includes independent variables, the main goals will likely be to test the statistical significance and characterize the relationship between the dependent and independent variables, that is, estimate the regression coefficients and associated confidence intervals or statistical tests. Guidelines for method selection for regression analyses are more complicated than those for detecting phylogenetic signal, so we discuss different scenarios separately.

9.5.2.1 Independent Variables Lack Phylogenetic Signal

When independent variables show no phylogenetic signal, all methods worked surprisingly well. In fact, there are few statistical problems introduced by phylogenetic signal in the residuals; even Logistf worked well, with little sign of inflated Type I errors. Therefore, it makes sense, before performing any other analysis, to first test for phylogenetic signal in the independent variables. If the estimates of phylogenetic signal are zero, then try Logistf or another logistic regression package (but use the Firth correction). It is necessary to recognize that by "no phylogenetic signal," we mean that the estimate of phylogenetic signal is exactly zero; absence of statistically significant phylogenetic signal is probably not a strict enough requirement to ignore phylogenetic signal all together, because the power of statistical tests for signal is often low (Blomberg et al. 2003).

9.5.2.2 Testing for Significant Regression Coefficients

If the goal of the analysis is to test the null hypothesis H_0: $b_1 = 0$, then RegOU or similar methods that ignore the binary nature of the dependent variable give a simple approach that in our simulations was remarkably statistically robust. This is not the heresy it might initially seem. The central limit theorem is remarkable: Given enough samples, the sum of their values will approach a Gaussian distribution, regardless of the distribution of any single sample. In principle, ignoring the unavoidable variation caused by Y taking only values of zero and one, RegOU should suffer loss of power compared to methods that account for this variation, yet we did not see a large loss of power in simulations with $N = 64$ (Fig. 9.4). A major disadvantage of RegOU, however, is that it does not give a model that fits the data. Therefore, it is hard to interpret the meaning of the regression coefficients, and it is not possible to simulate data from the model that has the statistical properties of the original data. This makes parametric bootstrapping impossible.

9.5.2.3 Fitting a Binary Regression Model

PLogReg, PGLMM, and MCMCglmm all performed reasonable well in fitting a model to data, although all had worrying defects. In particular, using the asymptotic approximations to the standard errors of the parameters, PLogReg tended to give inflated Type I errors for larger samples sizes, whereas MCMCglmm gave inflated Type I errors for smaller samples. Furthermore, all three methods showed lower-than-expected Type I errors under some situations, implying loss of statistical power (Fig. 9.3).

A solution to the Type I errors for PLogReg and PGLMM is to bootstrap the statistical tests for H_0:$b_1 = 0$. Bootstrapping guarantees against inflated Type I errors. Furthermore, bootstrapping the estimates of b_1 will give the most accurate confidence intervals and also show possible bias in the parameter estimates. Similarly, investigation of the joint distribution of parameter estimates in MCMCglmm will show possible problems with parameter estimation. Thus, although these methods have potential problems, those problems can be identified, and bootstrapping for PLogReg and PGLMM can solve the problem of inflated Type I errors. These two methods, used with parametric bootstrapping, provide the only approaches to fitting an appropriate model to phylogenetic binary data that have yet been validated.

9.6 Summary and Future Directions

Analyzing binary comparative data while accounting for potential phylogenetic correlations is not straightforward, but we hope our rough guidelines will help. It is important at the onset of analyses, however, to carefully consider your data. If you

have, for example, $N = 16$ taxa of which only three show one of the two possible trait states, then you should probably not embark on an analysis at all, particularly if the three taxa are closely related; the amount of information available in the data is unlikely to yield trustworthy statistical results. Before analyzing the dependent variable, test for phylogenetic signal in the independent variables; strong phylogenetic signal should suggest being particularly cautious in a regression analysis. Similarly, check for collinearity among the independent variables and remove those that seem unlikely to be informative on a priori grounds. Visually inspect the data: Are there observable patterns in the dependent variable? Do model fits look consistent with the data (e.g., Fig. 9.5)? If a statistical test tells you that a pattern is significant, then you should be able to see it in the data. After initial analyses, perform diagnostics by bootstrapping or by examining the parameter distributions from MCMCglmm. Use multiple methods, hoping that they give similar results; if they do not, then try to figure out why and which (if any) to trust. Finally, do not hesitate to use a model that you know is not strictly appropriate—phylogenetic regression for continuous traits, such as RegOU—as it still might give results that are statistically robust for hypothesis testing.

Phylogenetic models for the analysis of binary dependence variables need further statistical development. A simple, fast, robust method that does not suffer inflated Type I errors is still not available. Improvements in algorithms (e.g., **phylolm** {R}) will help. Another possible avenue is to pursue probit models; for MCMCglmm, the probit model had better statistical properties than the logit model, although probit models are not available for frequentist methods. A limitation of the three available methods we investigated is the absence of a true ML or REML score that can be used for model selection. Practically, this is not a huge hindrance, because all of the models can be used with classical backward or forward stepwise regression using, for example, P-values to decide which terms to include. Nonetheless, the popularity of backward and forward selection procedures has waned, and most researchers will now want an AIC score or a related information-theoretic metric. Finally, we have only investigated methods for binary data, but the same issues (although probably less severe) will appear for Poisson and other non-Gaussian data. Lots of problems are yet to be solved.

References

Blomberg SP, Garland T Jr, Ives AR (2003) Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution 57:717–745

Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, White JSS (2009) Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol Evol 24(3):127–135. doi:10.1016/j.tree.2008.10.008

Bonine KE, Gleeson TT, Garland T Jr (2005) Muscle fibre-type variation in lizards (Squamata) and phylogenetic reconstruction of hypothesized ancestral states. J Exp Biol 208:4529–4547

Brashares JS, Garland T Jr, Arcese P (2000) Phylogenetic analysis of coadaptation in behavior, diet, and body size in the African antelope. Behav Ecol 11(4):452–463

Diaz-Uriarte R, Garland T Jr (1996) Testing hypotheses of correlated evolution using phylogenetically independent contrasts: sensitivity to deviations from Brownian motion. Syst Biol 45(1):27–47

Diggle P, Heagerty P, Liang K, Zeger S (2004) Analysis of longitudinal data, 2nd edn. Oxford University Press, Oxford

Dlugosz EM, Chappell MA, Meek TH, Szafrañska P, Zub K, Konarzewski M, Jones JH, Bicudo JEPW, Careau V, Garland T Jr (2013) Phylogenetic analysis of mammalian maximal oxygen consumption during exercise. J Exp Biol 216:4712–4721

Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman and Hall, New York

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

Felsenstein J (1988) Phylogenies and quantitative characters. Annu Rev Ecol Syst 19:445–471

Felsenstein J (2012) A comparative method for both discrete and continuous characters using the threshold model. Am Nat 179(2):145–156. doi:10.1086/663681

Firth D (1993) Bias reduction of maximum likelihood estimates. Biometrika 80(1):27–38

Freckleton RP, Harvey PH, Pagel M (2002) Phylogenetic analysis and comparative data: a test and review of evidence. Am Nat 160:712–726

Garland T Jr, Dickerman AW, Janis CM, Jones JA (1993) Phylogenetic analysis of covariance by computer-simulation. Syst Biol 42(3):265–292

Garland T Jr, Harvey PH, Ives AR (1992) Procedures for the analysis of comparative data using phylogenetically independent contrasts. Syst Biol 41:18–32

Garland T Jr, Midford PE, Ives AR (1999) An introduction to phylogenetically based statistical methods, with a new method for confidence intervals on ancestral values. Am Zool 39:374–388

Gelman A, Carlin JB, Stern HS, Rubin DB (1995) Bayesian data analysis, 1st edn. Chapman and Hall, London

Gelman A, Hill J (2007) Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, New York

Grafen A (1989) The phylogenetic regression. Trans R Soc Lond B, Biol Sci 326:119–157

Hadfield JD (2010) MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. J Stat Softw 33:1–22

Hadfield JD (2012) MCMCglmm course notes. http://cran.r-project.org/web/packages/MCMCglmm/vignettes/CourseNotes.pdf

Hadfield JD, Nakagawa S (2010) General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. J Evol Biol 23(3):494–508. doi:10.1111/j.1420-9101.2009.01915.x

Hansen TF (1997) Stabilizing selection and the comparative analysis of adaptation. Evolution 51:1341–1351

Hansen TF, Orzack SH (2005) Assessing current adaptive and phylogenetic inertia explanations of trait evolution: the need for controlled comparisons. Evolution 59:2063–2072

Harvey PH, Pagel MD (1991) The comparative method in evolutionary biology. Oxford University Press, Oxford

Heinze G, Ploner M, Dunkler D, Southworth H (2013) logistf: Firth's bias reduced logistic regression. Vol R package version 1.21

Heinze G, Schemper M (2002) A solution to the problem of separation in logistic regression. Stat Med 21:2409–2419

Ho LST, Ane C (2014) A linear-time algorithm for gaussian and non-gaussian trait evolution models. Systematic Biology in press

Ives AR, Garland T (2010) Phylogenetic logistic regression for binary dependent variables. Syst Biol 59(1):9–26. doi:10.1093/sysbio/syp074

Ives AR, Helmus MR (2011) Generalized linear mixed models for phylogenetic analyses of community structure. Ecol Monogr 81:511–525

Jarman PJ (1974) The social organisation of antelope in relation to their ecology. Behaviour 48:215–267

Judge GG, Griffiths WE, Hill RC, Lutkepohl H, Lee T-C (1985) The theory and practice of econometrics, 2nd edn. Wiley, New York

Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP, Webb CO (2010) Picante: R tools for integrating phylogenies and ecology. Bioinformatics 26:1463–1464

Lavin SR, Karasov WH, Ives AR, Middleton KM, Garland T Jr (2008) Morphometrics of the avian small intestine, compared with non-flying mammals: a phylogenetic approach. Physiol Biochem Zool 81:526–550

Martins EP, Diniz JAF, Housworth EA (2002) Adaptive constraints and the phylogenetic comparative method: a computer simulation test. Evolution 56(1):1–13

Martins EP, Garland T Jr (1991) Phylogenetic analyses of the correlated evolution of continuous characters: a simulation study. Evolution 45:534–557

Martins EP, Hansen TF (1997) Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. Am Nat 149:646–667. Erratum 153:448

McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman and Hall, London

McCulloch CE, Searle SR, Neuhaus JM (2008) Generalized, linear, and mixed models. Wiley, Hoboken, NJ

Pagel M (1994) Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proc R Soc Lond Ser B Biol Sci 255(1342):37–45

Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290

Revell LJ (2010) Phylogenetic signal and linear regression on species data. Methods Ecol Evol 1:319–329

Revell LJ (2012) Analyzing continuous character evolution on a phylogeny. Integr Comp Biol 52:E145–E145

Revell LJ, Harmon LJ, Collar DC (2008) Phylogenetic signal, evolutionary process, and rate. Syst Biol 57(4):591–601. doi:10.1080/10635150802302427

Rezende EL, Diniz JAF (2012) Phylogenetic analyses: comparing species to infer adaptations and physiological mechanisms. Compr Physiol 2(1):639–674. doi:10.1002/cphy.c100079

Schluter D, Price T, Mooers AO, Ludwig D (1997) Likelihood of ancestor states in adaptive radiation. Evolution 51:1699–1711

Villemereuil P, Gimenez O, Doligez B (2013) Comparing parent-offspring regression with frequentist and Bayesian animal models to estimate heritability in wild populations: a simulation study for Gaussian and binary traits. Methods Ecol Evol 4(3):260–275. doi:10.1111/2041-210x.12011

Chapter 10
Keeping Yourself Updated: Bayesian Approaches in Phylogenetic Comparative Methods with a Focus on Markov Chain Models of Discrete Character Evolution

Thomas E. Currie and Andrew Meade

Abstract   Bayesian inference involves altering our beliefs about the probability of events occurring as we gain more information. It is a sensible and intuitive approach that forms the basis of the kinds of decisions we make in everyday life. In this chapter, we examine how phylogenetic comparative methods are performed within a Bayesian framework, introducing some of the main concepts involved in Bayesian statistics, such as prior and posterior distributions. Many traits of biological and evolutionary interest can be modelled as being categorical, or discretely distributed, and here, we discuss approaches to investigating the evolution of such characters over phylogenetic trees. We focus on Markov chain models of discrete character evolution and how these models can be assessed using maximum-likelihood and Markov Chain Monte Carlo techniques of parameter estimation. We demonstrate how this can be used to test functional hypotheses by examining the correlated evolution of different traits, illustrated with examples of sexual selection in primates and cichlid fish. We show how the order of trait evolution can be determined (potentially providing a stronger test of causal hypotheses) and how competing hypotheses can be assessed using Bayes factors. Attractive features of these Bayesian methods are their ability to incorporate uncertainty about the phylogenetic relationships between species and their representation of results as probability distributions rather than point estimates. We argue that Bayesian methods provide a more realistic way of assessing evidence and ultimately a more intellectually satisfying approach to investigating the diversity of life.

T. E. Currie (✉)
Centre for Ecology and Conservation, Biosciences, College of Life and Environmental Sciences, University of Exeter, Penryn Campus, Penryn, Cornwall TR10 9FE, UK
e-mail: t.currie@exeter.ac.uk

A. Meade
School of Biological Sciences, University of Reading, Whiteknights Campus, Philip Lyle Building, Reading, Berkshire RG6, UK

10.1 Introduction

What are the chances of England winning the next football World Cup?[1] An initial estimate, for those lucky enough to be unfamiliar with the performances of the national team, might come from knowing that 32 teams take part in the finals of the tournament. Presuming that England qualify for this stage and prior to knowing anything else about England or football, we might assume that all teams have an equal chance of winning, meaning that the probability of England being victorious is 1/32. However, differences in abilities between teams make it unsafe to assume that all teams have an equal chance of winning. A look at how well England have done in previous tournaments may also serve as a useful guide. England are only one of eight teams to have won the competition[2] since it started in 1930. Therefore, we might adjust our estimate of the odds of success to 1:7. We might further look at their record in tournaments over the last 15 years (i.e. since 1998) and see that they have failed to make it beyond the quarter-finals and sometimes only as far as the round of 16. Given this information, we might alter our estimation of the chances of success down slightly. Once the World Cup finals kick off, we might also update these beliefs about England winning we had prior to the tournament based on how well (or more likely how badly) they play. If by some miracle they made it through to the final, then even the most pessimistic fan would have to adjust their estimate to somewhere closer to an even chance of success.

Altering our beliefs about the probability of events occurring as we gain more information is extremely sensible and obviously forms the basis of the kinds of decisions we make in everyday life. It is this kind of reasoning that forms the basis of Bayesian inference. In this chapter, we will examine how this is applied to phylogenetic comparative methods, with a particular focus on traits that are categorical or discretely distributed. After first introducing some of the concepts involved in Bayesian statistics, we will discuss earlier approaches to investigating the evolution of discrete characters over evolutionary trees and how Bayesian approaches can overcome some of the limitations of these approaches. In the Online Practical Material (hereafter OPM) available at http://www.mpcm-evolution.com, we will provide specific examples of how Bayesian phylogenetic comparative methods are used to investigate interesting evolutionary questions.

10.2 Bayesian Inference

The Bayesian approach to probability can be summarized as follows: We have an initial, or prior, belief about the probability of something being true which we adjust based on new information to arrive at our updated, or posterior, belief

---

[1] This way of introducing this topic owes a debt to Ronquist et al. (2009), although here we focus on the inadequacies of our national football team rather than the success of the Swedish ice hockey team.
[2] 1966.

(Ronquist et al. 2009). This approach to thinking about probabilities was formalized by Thomas Bayes in the eighteenth century in what is known as Bayes theorem, which states

$$Pr(B|A) = \frac{Pr(A|B)Pr(B)}{Pr(A)} \tag{10.1}$$

This formula is read as the posterior probability of B given A ($Pr(B|A)$) equals the probability of A given B (($Pr(A|B)$)), multiplied by the probability of B (Pr(B)) and divided by the probability of A (Pr(A)) (Link and Barker 2009).

To go back to our football example, by some miracle let us imagine that England have made it through to the World Cup final, and yet more miraculously, they have managed to score first. We can ask, what is the probability that England will go on to win the final? We can calculate this using Bayes theorem which will tell us the probability that England will win given the new information that we have about England scoring first.

$$P(\text{England win}|\text{Score first}) = \frac{P(\text{Score first}|\text{England win})P(\text{England win})}{P(\text{Score first})} \tag{10.2}$$

In order to work this out, we need to know the prior probability of England scoring first (i.e. in general, how common is it for England to be the team that scores first?), and also for those games where England do win, the probability that they scored first. Looking back over the past records, we find that England score first in around two-thirds of their games (i.e. a probability of 0.67), and in the games that they win, they score first 80 % of the time (i.e. a probability of 0.8). For the prior probability of England winning, let us also say that at the beginning of the game, there was an equal chance that either team could win.

$$P(\text{England win}|\text{Score first}) = \frac{0.8 \times 0.5}{0.67} \tag{10.3}$$

Working this through gives us a probability of around 0.6 that England having scored first will indeed end years of hurt and lift the World Cup trophy. In the next section, we will see how these concepts of using information to update our prior beliefs and arrive at our posterior beliefs are used in the context of PCMs.

10.3 Phylogenetic Comparative Methods and Discrete Characters

Comparing traits across species is a fundamental part of biology and enables us to test hypotheses about the functions of the traits and gain insights into their evolutionary history. In phylogenetic comparative methods (PCMs), we map traits of

interest onto a phylogeny (which shows how the species are related evolutionarily) and work backwards to make inferences about the pattern and process of change in these traits. Evolutionary trees contain information that enables us to examine biological diversity and make inferences about where, when, and how traits have changed over time and to test hypotheses about why such diversity exists (Pagel 1999). As the chapters in this book demonstrate, a variety of methods have been developed that can address a number of important evolutionary questions.

PCMs can be used to analyse both continuously and discretely distributed data. Typical continuous characters involve measurable features of size and distance (e.g. body size, wing length), while discrete traits are those that can be thought of as falling into distinct categories (e.g. mating system, feeding behaviour) or reflecting presence or absence of a certain trait.[3] The majority of the other chapters in this book deal with methods designed for continuous data (but see Chap. 9 by Ives and Garland, Chap. 11 by de Villemereuil and Nakagawa, and Chap. 16 by Beaulieu and O'Meara). Therefore, here, we will deal primarily with discrete characters and how they can be implemented in a Bayesian framework. At the end of this chapter, we will briefly see how this approach can be extended easily to methods that utilize continuous traits. In order to understand the benefits a Bayesian approach can have, we will first examine some of the other methods that have been developed to analyse discrete traits.

10.3.1 Modelling the Evolution of Discrete Characters

10.3.1.1 Parsimony

The earliest PCMs for discrete traits were based on the idea of maximum parsimony, i.e. minimizing the number of evolutionary changes (Maddison 1990). Given a certain distribution of character states at the tips of the tree, there are many different possible ways that a character can change over the tree. Parsimony methods find the pattern[4] that involves the lowest number of transformations between character states. Parsimony works under an implicit assumption that the rate of evolution is slow,[5] and when this is the case, it leads to fairly accurate reconstructions of character evolution (Huelsenbeck et al. 2003). In the basic implementation of parsimony, only a single change can occur along a branch, and a change from one state to any other is equally probable. This approach can be modified slightly by proposing a cost matrix, wherein certain changes incur a higher tariff (Maddison and Maddison 2009).

---

[3] It should be noted that count data, such as clutch size, are also technically discrete yet are not categorical.

[4] or pattern*s* if more than one solution is possible.

[5] relative to rate at which new lineages form.

10.3.1.2 Maximum Likelihood

Parsimony is somewhat limited in that it only provides reconstructions of the pattern of changes in the character of interest. An alternative is to adopt a model-based approach that seeks to make inferences about the process of character evolution (from which likely patterns of changes can be reconstructed) (Pagel 1994a; Sanderson 1993). For example, the evolution of a binary trait (i.e. taking values of 0 and 1) can be modelled in a very simple manner invoking just two parameters, a rate of change from 0 to 1 and a rate of change from 1 to 0. In fact, we can even propose a simpler model with just one parameter, if we assume that these two rates are equal. We can use the information about the distribution of 0s and 1s at the tips of the tree, and the branch lengths of the tree, in order to estimate these rates of change. A slow rate of change is likely to lead to outcomes where closely related species generally exhibit the same character, while under faster rates of change, even closely related species will not necessarily share similar characters (Fig. 10.1).

In practical terms, character evolution is modelled using a continuous-time Markov chain, which is a mathematical system that transitions at random between different character states. One of the features of a Markov process is that it has no "memory"; the probability of change from one state to another depends only on the current state and not on what has happened previously. A transition matrix is used to describe the rate of change between different character states (Fig. 10.1).[6] These rates are known as instantaneous rates of change and reflect the probability of change over an infinitesimally small amount of time. This approach is extremely flexible and can be readily extended to multiple character states, and more complexity can be added to reflect variation in the rate of change between different character states. For a given number of character states, different models, which represent alternative evolutionary hypotheses, can be constructed and assessed as to how well they explain the observed data. By setting certain rates of change between character states to be zero, hypotheses about different evolutionary pathways can be tested (Currie et al. 2010; Hibbett 2004; Pagel 1994a). Using the example in Fig. 10.1, if we wanted to test the hypothesis that a certain can be gained but never lost, then we could set q10 to 0, indicating that the transition from 1 to 0 cannot occur. This could then be tested against other models, which allow changes in both directions. Using this approach, Currie et al. (2010) examined competing hypotheses relating to alternative evolution pathways of human political organization and found that changes follow incremental steps of increasing hierarchical complexity, with larger jumps not occurring.

The fit of the models to the data and the value of the parameters can be assessed by calculating what is known as the likelihood function. Using maximum-likelihood (ML) estimation, we search for the values of the model of evolution that give the best description of the data (i.e. the values that maximize the likelihood function)

---

[6] For more on transition matrices, see Chap. 16 by Beaulieu and O'Meara in this volume.
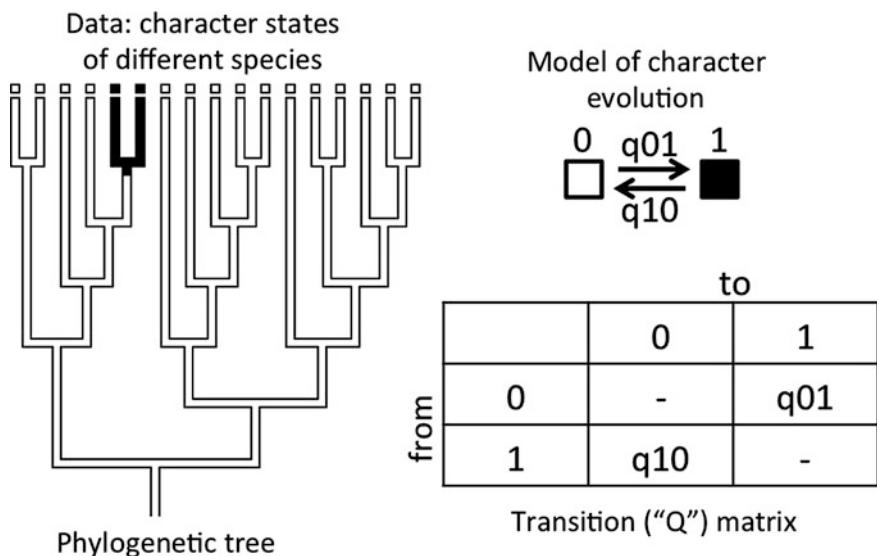
Fig. 10.1 Model-based phylogenetic comparative methods use character data mapped onto the tips of a phylogenetic tree to infer the parameters of a model of character evolution. The evolution of binary, or two-state, character can be modelled simply with two parameters, a rate of change from 0 to 1 and rate of change from 1 to 0. In this example, closely related species tend to possess the same character state, indicating a slow rate of change. When using continuous-time Markov chain methods, these rate parameters are represented in a transition (or "Q") matrix

(Pagel 1999). Using the example in Fig. 10.1, let us imagine we are estimating a simple one-parameter model (i.e. q01 = q10). If we were to propose a high value for this rate parameter, it would produce a low likelihood. Proposing a lower value would result in a higher (i.e. better) likelihood. Proceeding like this, we would propose lower values until we reached a point where lower values began producing lower (i.e. worse) likelihoods (see Fig 10.2). By comparing the maximum likelihoods of different models, we can test between different hypotheses about the evolutionary process that gave rise to the observed data (see below, and Garamszegi and Mundry Chap. 12, this volume).

One of the key advantages of model-based approaches over parsimony is that they use the information about the branch lengths of the tree in the analyses. For example, because more changes are likely over a long time period than a short one, we would intuitively expect that saying something about the likely character state in ancestral species becomes more difficult the further back in time you go. However, parsimony discards this information and attaches the same probability to the ancestor of two sister species, regardless of whether they diverged 100 years ago or 100 million years ago. Under model-based methods, however, the probability of the inferred ancestral state is affected by the branch lengths of the tree (see Fig. 10.3).
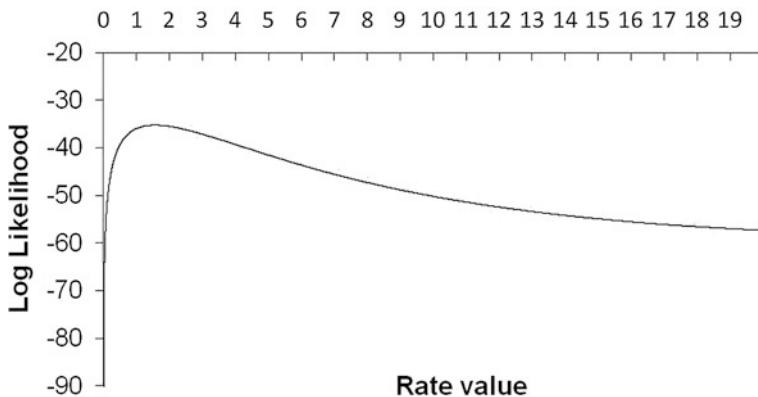
Fig. 10.2 Example of a likelihood surface showing how the likelihood of the model changes with different values of the rate parameter. In this example, the value of the parameter that maximizes the likelihood is around 1.6

10.4 Bayesian Methods

10.4.1 Dealing with an Uncertain World

Under maximum-likelihood methods, we find point estimates of parameter values that provide the best fit to the data. However, there may be a range of parameter values (which give slightly lower likelihoods) that still provide a reasonable description of the data. Rather than just calculating the maximum-likelihood estimate of parameter values, we can employ a Bayesian approach to estimate the *posterior probability distribution* of values (Fig. 10.4) (Pagel and Meade 2005). In other words, we start with some prior distribution reflecting the possible values of these parameters (see below); we then update this distribution based on the analysis to arrive at the posterior distribution of the parameter values given the observed data. This allows us to incorporate uncertainty in the parameter estimates into our analyses.

This framework also allows us to incorporate *phylogenetic uncertainty* into our analyses (Huelsenbeck et al. 2000; Pagel et al. 2004; Huelsenbeck and Rannala 2003; and Chap. 3 this volume). Rarely can we know the phylogenetic relationships between species without error. Importantly, our inferences about how traits have evolved may differ, depending on what we assume about the phylogenetic relationships between species. Rather than attempting to represent these phylogenetic relationships with a single tree, it is more principled to use a collection of trees that represent likely alternative hypotheses about how species are related. Furthermore, de Villemereuil et al. (2012) demonstrate that using a Bayesian approach to incorporate phylogenetic uncertainty in analyses of linear models is more accurate (i.e. reduces the error rate associated with estimates of model parameters) than
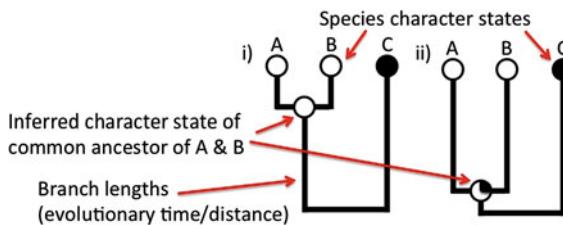
Fig. 10.3 Model-based phylogenetic comparative methods make use of the information from the branch lengths of phylogenetic trees. Here, we can see how knowledge of the branch lengths (which represent evolutionary time, or more generally evolutionary distance, if the branch lengths are not in units of time) can affect estimation of the ancestral character states. In situation (*i*), species *A* and *B* diverged relatively recently, meaning we can be more certain that their common ancestor shared the same character state. In situation (*ii*), however, divergence occurred further in the past, which means there is more uncertainty about the reconstruction of this character. In this hypothetical example, the analyses indicate that the probability that the common ancestor had the same character state as species *A* and *B* is around 0.75. Parsimony analyses do not use branch-length information and so would return the same answer under both scenarios
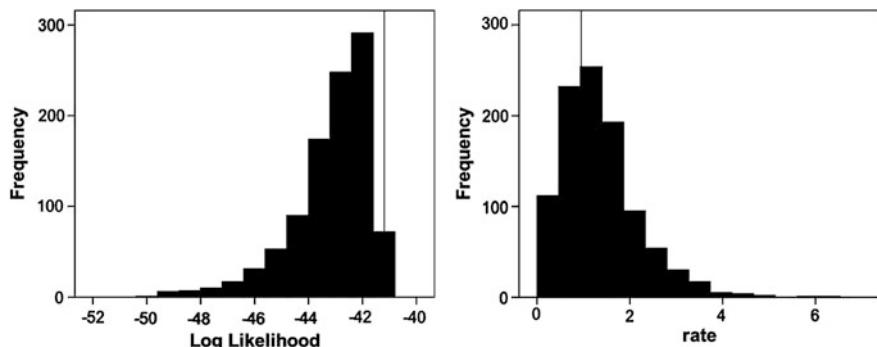
Fig. 10.4 Bayesian MCMC methods produce a posterior distribution of likelihoods (*left*) and parameter values (*right*). The straight vertical lines represent the equivalent estimates under a maximum-likelihood analysis. Notice in the Bayesian analysis, the likelihoods reach the maximum likelihood but do not exceed it (that is what maximum means!). The rate parameter estimates fall either side of the maximum-likelihood value as slightly higher or lower values produce slightly worse likelihoods

running a regular PGLS[7] analysis using a single tree. The collection of trees could represent different published hypotheses, but probably more commonly comes from a posterior sample of phylogenetic trees from a Bayesian method of phylogenetic inference (i.e. trees created from an analysis of genetic or morphological data, where phylogenetic trees are sampled in proportion to their probability). For example, Bayesian posterior samples of 10,000 phylogenetic trees representing the

---

[7] Phylogenetic generalized least squares.

evolutionary relationships between (respectively), (i) primates, (ii) carnivorans, (iii) even-toed ungulates and cetaceans, and (iv) odd-toed ungulates, are available from the 10kTrees website[8] with the intention of their being employed in comparative analyses (Arnold et al. 2010). With Bayesian PCMs, we can naturally incorporate phylogenetic uncertainty by performing the analysis over a sample of trees.[9] In the terms of Bayes theorem we introduced earlier essentially, we are asking, what is the probability of the model of evolution given the data and the sample of phylogenetic trees? To do this, we need to give values for the prior probabilities of (i) the data given the model and the trees, (ii) the model, and (iii) the trees.

Incorporating these different forms of uncertainty into our analyses in this way is an intellectually satisfying (but practically challenging) feature of Bayesian PCMs because it gives a better idea about how strong the support for any particular hypothesis actually is (as good scientists, we should always be sceptical about our models and look to see how robust they are to different assumptions). For example, if there is a strong signal in the data, then the posterior distribution of parameter values should cluster closely around the maximum-likelihood estimate. However, if there is a weaker signal, the posterior will be wider and may be more likely to overlap with alternative models. Likewise, we need to assess whether the strength of support for a particular hypothesis is affected heavily by the particular evolutionary relationships given by different phylogenetic trees.

10.4.2 MCMC Estimation

In order to estimate the posterior distribution of parameter values, we can use the Markov chain Monte Carlo (MCMC) procedure to explore and take a sample of values from "parameter space". In this technique, parameter values are sampled in proportion to their posterior probability (i.e. more probable values are sampled more frequently). Essentially, the chain starts with some approximate parameter values and a tree drawn at random from the tree sample. The likelihood of the data given these values and the tree is then calculated. At the next step, the values of the parameters and the tree from the sample are changed at random and the likelihood is again calculated. The new parameter values are either accepted, or the old values are retained. If the new likelihood is an improvement on the previous likelihood, then the new values are accepted. Otherwise, they are accepted only with a certain probability, depending on how worse the new likelihood is.[10] This process is repeated many times. Eventually, the chain ends up searching more often through

---

[8] http://10ktrees.fas.harvard.edu/

[9] See Garamszegi and Mundry Chap. 12, this volume, for an example of how to incorporate phylogenetic uncertainty within an Information Criterion framework.

[10] The process described here relates to the Metropolis–Hastings MCMC algorithm. However, other algorithms such as the Gibbs sampler are also available that follow different rules about how they accept new values and explore the posterior distribution.
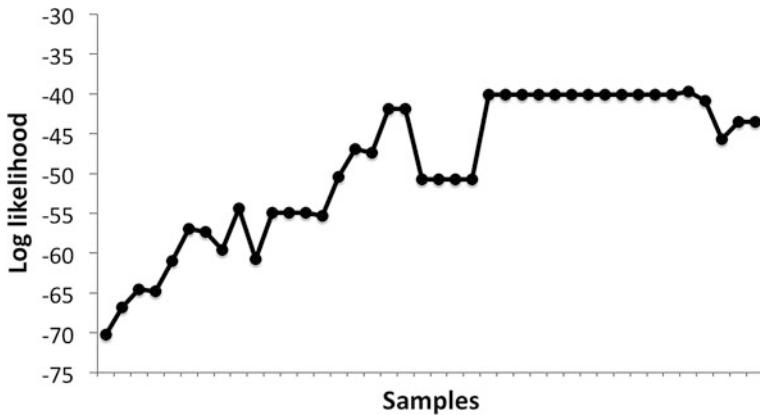
Fig. 10.5 Example of the early stages of an MCMC sampling procedure. From a random starting position, the algorithm gradually finds parameter values that provide a better fit to the data, until it reaches the region of the posterior distribution. The initial phase where the likelihoods are generally increasing is known as the burn-in

the areas of parameter space that provide the highest likelihoods (Fig. 10.5). At this point, which is known as *convergence*, the chain is sampling from the posterior distribution.

There are two important issues we need to be aware of in evaluating the output of MCMC analyses. Firstly, at the beginning, the MCMC has generally started at a point away from the posterior distribution and the initial stages are characterized by a "hill-climbing" phase as the chain moves from the low-likelihood region of parameter space (Fig. 10.5). The chain is described as converging on the posterior distribution. As we are looking to estimate the posterior distribution, we need to discard the initial pre-convergence, or "burn-in", phase of the MCMC. Secondly, the way new values in the MCMC procedure are accepted or rejected means successive steps in the chain may be correlated with one another. Potentially, this *autocorrelation* can lead to a biased sample that is not representative of the posterior distribution. While running the analysis for a sufficiently long period of time would ameliorate this problem, this could lead to a large and unnecessarily unwieldy amount of output. The usual solution, known as *thinning*, is to take samples from the MCMC at regular intervals (e.g. every 100 steps), rather than output every single iteration of the chain. The sampling frequency can be determined by examining the degree of autocorrelation from initial exploratory analyses.

10.4.3 Priors

In order to come up with our posterior beliefs, Bayesian approaches require us to specify the prior beliefs. For these PCMs, this means specifying what values we
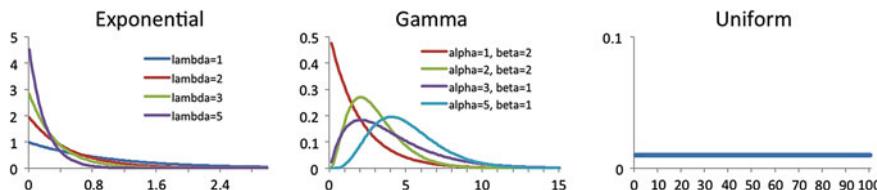
Fig. 10.6 Examples of different prior probability distributions that can be used in Bayesian phylogenetic comparative analyses. Under a uniform distribution, all values within a given range are equally probable. Lower values are more probable under exponential distributions (here shown with varying distribution parameters, *lambda*). Gamma distributions have two parameters (shape: *alpha* and scale: *beta*), which can give rise to a variety of different shaped distributions. Under certain parameter values, the gamma distribution is similar to an exponential, and others give a humped distribution with moderate values being most probable

think characterize the rates of evolution. Our parameter values are continuous and in theory could take an infinite possible number of states. To make the process tractable, we instead specify a prior distribution for the parameters (Ronquist et al. 2009). This is an extremely important aspect of Bayesian methods and is both a blessing and a curse. One advantage is that it allows us to incorporate other sources of knowledge so that we do not waste time exploring answers that cannot possibly be true. The downside is that often we do not have much information about what these priors should be. In the case of PCMs, it is difficult to know what values of the rates of evolution are sensible a priori. Caution needs to be exercised as if there is not a strong signal in the data, then the particular prior used can have a big effect on the posterior and therefore may determine the answer we arrive at (Pagel and Meade 2005). Indeed, the weaker the signal, the more the posteriors sample the priors, meaning that in the extreme, the posterior sample will be the same as the prior.

There are a number of different prior distributions that can be used (Fig. 10.6). The simplest and least restrictive is the uniform distribution, which assumes that any value between two specified points is equally likely. An alternative is the exponential distribution, which assumes that lower values are more likely (which may be plausible biologically if we share the assumption of parsimony that rates of evolution are generally low). Another common prior is the gamma distribution, which can take a variety of shapes, in some cases approximating an exponential, in other cases assuming that mid-range values are most likely (this may be preferable if we know certain changes have definitely occurred and therefore rate values must be greater than zero). In cases where there is not much information to guide our choices, it is preferable to use a uniform distribution, as this has the fewest assumptions. However, if the signal in the data is not strong enough, it may be necessary to specify a stronger prior. This can be assessed by examining the posterior distributions of the parameters and seeing whether they centre around a particular value or whether they have maintained a relatively flat distribution. Different prior distributions can be explored, and their effect on the results can be

examined. If a particular result is determined by the choice of prior, then there should be good, justifiable reasons for choosing that prior. The results of maximum-likelihood analyses are a useful way of guiding the choice of prior values as they can give an indication of the mid-point of the posterior distribution (although they will not define the range of the prior). As a general rule of thumb if the posterior distribution appears to be truncated at either the upper or lower limits defined by an informative (i.e. non-uniform) prior, then the limits of the prior should be adjusted. One useful approach is to use *hyperpriors*, where the values of the prior distribution are not set but can also vary (see glossary and the OPM). This provides more information than a uniform distribution, but is less constraining than a regular, single prior distribution.

10.5 Assessing Models of Evolution

When using model-based methods with discrete characters, we are generally not interested in the values of the rate parameters themselves.[11] Instead, the aim is to compare different models. Usually, we are more interested in whether one model (which represents a particular hypothesis) is a better explanation of the data than another model (representing an alternative hypothesis). For example, we might want to know whether change between two states of a binary trait occurs at the same rate in both directions or whether changes from 0 to 1 occur at a higher rate than from 1 to 0. We will see later how this approach is used to assess whether two traits evolve together in a correlated fashion or whether they have evolved independently of each other.

Under maximum likelihood, if one model can be thought of as "nested" within another model (i.e. one model is a simpler version of another model, with certain parameters set to be equal to each other, or "switched off" by setting them to zero), then models can be compared via a likelihood ratio test (Posada 2009). The likelihood ratio statistic is calculated as double the difference of the log-likelihoods of the simpler and the more complex model. Generally, this statistic is assumed to approximate a chi-squared distribution, with degrees of freedom determined by the difference between the number of parameters in each model. This therefore takes into account the fact that a nested model with more parameters will never produce a worse maximum likelihood, and it is easier to get a higher likelihood with more parameters. A more general framework for comparing both nested and non-nested models is model-selection procedures involving information criteria (e.g. AIC, BIC), which contain an explicit term that takes into account the number of parameters in a model (models with too many parameters get "penalized") (Burnham and Anderson 2002; see also Garamszegi and Mundry Chap. 12, this volume).

---

[11] They are a function of the data, model, and distribution of tree used in the analysis, which makes them hard to compare across analyses.

Exactly the same idea of comparing alternative models is at the heart of Bayesian approaches to statistical inference. Whereas under maximum likelihood we are comparing the likelihoods of two models in a Bayesian framework, we need to compare two posterior probability distributions, using what is known as the marginal likelihood of each model. The marginal likelihood of a model is its likelihood scaled by the prior probabilities and integrated over all the trees in the sample and all values of the rate parameters (Pagel and Meade 2005).[12] This incorporation of the priors means that a model with more parameters is not unduly favoured (just as extra parameters are penalized in the likelihood ratio test and AIC methods described above). In a Bayesian framework, we can test between different hypotheses by calculating a measure known as a Bayes factor (Kass and Raftery 1995).[13] This is effectively a ratio of probabilities of the data given the different hypotheses and can be calculated as the ratio of the marginal likelihoods of models. Bayes factors are interpreted somewhat subjectively with rules of thumb being employed to assess the strength of evidence in favour of one hypothesis or another. Since these values are based on log-likelihoods, the difference between them can be doubled so that they are on the same scale as likelihood ratio statistics. A value of less than zero is obviously evidence against the main hypothesis (and therefore evidence in favour of the alternative hypothesis). According to Kass and Raftery (1995), values between 0 and 2 are only just in favour of the hypothesis and not worth placing too much confidence in, 2–6 are described as "positive" evidence, 6–10 are "strong", while more than 10 is "very strong". In the following sections, we will see how these concepts are applied in some examples of comparative analyses using Bayesian methods.

10.6 Using Model-Based Methods to Test Functional Hypotheses

Most functional hypotheses take the form of arguing that a certain trait reflects an adaptive response to some other variable. This underlies the classic textbook examples of natural selection, variation in beak shapes in finches being a response to different diets, and changes in the frequencies of black and white variants of the peppered moth being linked to changing environments caused by industrialization. We can compare a range of different species to ascertain whether a general relationship between our variables of interest exists. This comparative method is one

---

[12] An approximation of the marginal likelihood is part of the output of the program used in the practical section that accompanies this chapter.

[13] Note that using Bayes factors (and model selection criteria such as AIC), it is possible to find evidence *for* a null hypothesis, something that is not possible in classical, frequentist statistics where the null hypothesis can only be rejected.

of the three ways that evolutionary or behavioural ecologists assess evidence for adaptive hypotheses (the others being experiments and optimality modelling) (Davies et al. 2012).

Pagel (1994a) developed an elegant method for testing whether two binary traits have indeed evolved together. This method is implemented in the program BayesTraits[14] and enables both maximum-likelihood and Bayesian MCMC methods of estimation. To illustrate this approach, let us consider the hypothesis that oestrus advertisement by females reflects an adaptation to living in multimale groups (Pagel 1994b; Domb and Pagel 2001) (we will return to this example later in the chapter and in the OPM). Prominent sexual swellings in female primates indicate when a female is fertile and most likely to conceive. They are hypothesized to have evolved in species that have groups with multiple males, who compete for access to the females. Looking at raw data from 60 species of Old World monkeys and apes, there appears to be an association between these two variables. Nineteen species have both multimale groups *and* females with conspicuous oestrus, 28 species lack either trait, nine species have multimale groups but lack oestrus advertisement, while none of the species have only single-male groups with females who advertise.[15] Of these 60 species, many are macaques (almost all of which possess both traits) and many are gibbons (all of which lack these traits). There is a strong possibility that we might be overestimating the strength of the association between these traits unless we adequately incorporate the phylogenetic relationships between these species.

Under Pagel's method, we explicitly compare a model of evolution in which these two traits evolve together (a dependent, or co-evolutionary model) and model in which the two traits evolve without affecting each other at all (an independent model). For two binary traits, there are four possible ways that these traits can co-occur across species (i.e. 00, 01, 10, 11). The dependent model proposes that there are 8 possible ways that traits can change between these four possible states (Fig. 10.7). For example, a species with a single-male system and without oestrus advertisement can develop either a multimale system or oestrus advertisement (it is unlikely that both traits change at exactly the same time, and this cannot happen under this model). The instantaneous rates of these changes make up the parameters of our model of evolution (i.e. the Q matrix we came across earlier). Under this model, the rate of change of one trait depends on the state of the other. It is this characteristic that makes this a model of co-evolution. It follows from this that an independent model would be one where the rate of change of one trait does not depend on the state of the other. Such a model can be achieved if we set the relevant parameters to be equal to each other (e.g. if the rate of change from single- to multimale systems is the same regardless of whether oestrus advertisement is present or not) (Fig. 10.7).

---

[14] http://www.evolution.rdg.ac.uk/BayesTraits.html

[15] Four species lack data for one of the traits, which illustrates that these methods can handle missing data; the likelihood is simply integrated over all possible character states in these cases. See also chap. 11 by de Villemereuil and Nakagawa for a discussion of the issues surrounding missing data and how to deal with them.
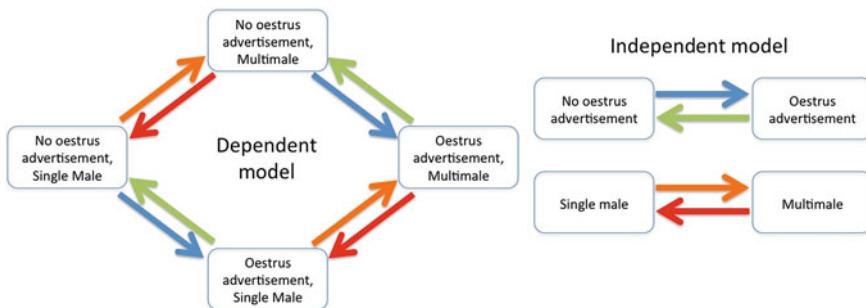
Fig. 10.7 Dependent and independent models of evolution under Pagel's method ([1994a](#)). The independent model can be seen to be a simpler version of the dependent model, which occurs when the arrows diagonally opposite each other (i.e. those that are the same colour) have the same rate of evolution

We can run analyses using Old World monkey and ape data over a sample of phylogenetic trees, which represent the uncertainty about the phylogenetic relationships between these species. Figure 10.8 shows how the MCMC algorithm samples from the posterior distribution. This is done for both the dependent and the independent models of evolution. Thus, for our primate data, we obtain estimates from these preliminary analyses of the marginal mean of ~ 41 log-units for the dependent model and ~ 46 log-units for the independent model. Comparing the dependent model to the independent model for our primate data gives us a Bayes factor[16] of 10, which means there is strong evidence that oestrus advertisement has indeed co-evolved with multimale systems. Therefore, the association between the traits we saw when we examined the raw data is not simply an artefact of the historical relationships between the species.

This example shows how we can get some measure of whether two traits have evolved together or not. We can also use these methods to investigate more about the specific evolutionary history of the traits we are interested in and build up a picture of where and when they have changed. A neat thing about this approach is that we can go beyond simple measures of association and examine the order in which the traits have occurred, particularly if we explicitly reconstruct the likely state of traits at ancestral nodes in the tree. This can provide a much stronger test of causal hypotheses. If changes in a certain trait are hypothesized to precede changes in another trait, and if our analyses indicate that the order of trait changes is actually the other way round, then we can reject that hypothesis.[17] We will return to our primate example in the OPM to give a demonstration of how this is done.

---

[16] on the scale discussed earlier.

[17] It is important to note that while we can falsify causal hypotheses in this way, if we do find evidence for the hypothesized order of trait changes, this does not prove causation but is at least consistent.
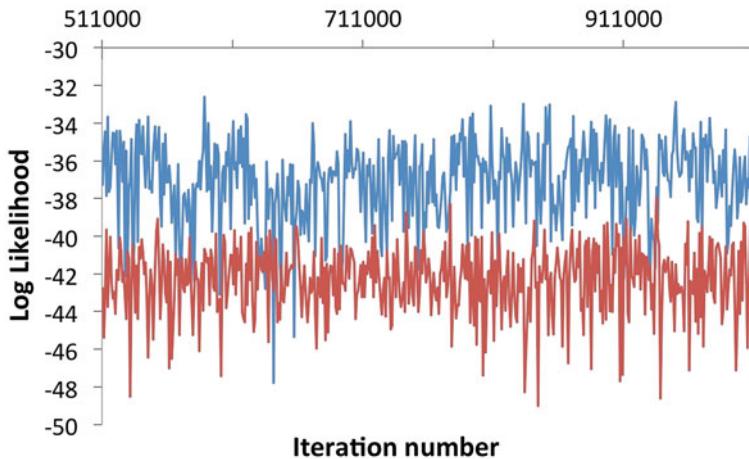
Fig. 10.8 Example of MCMC sampling from the posterior distribution under dependent (*blue*) and independent (*red*) models of evolution. The dependent model is generally returning higher likelihoods, indicating that it is a better fit to the data than the independent model. Note that the initial burn-in phase of the chain has been discarded

A nice example of this approach is provided by Fitzpatrick et al. (2009), who show how female promiscuity, which creates increased sperm competition, leads to larger and faster sperm. The authors examined the strength of sperm competition (based on breeding characteristics) and the speed and size of spermatozoa in 29 species of cichlid fish from Lake Tanganyika in eastern Africa. Ancestral state reconstruction using the Markov chain approach we have been discussing indicated that it was likely that the common ancestor of all these species experienced low sperm competition and had small, slow sperm. Examining the rates of change in their dependent models of evolution, they found that some parameters frequently took a value of zero, providing evidence that this transition had not occurred during the evolutionary history of these species. This allowed them to assess the likely order in which these traits changed. Their analyses indicate that sperm initially became faster before getting bigger, and importantly, both sperm size and speed increase *after* increases in levels of sperm competition, which is consistent with the idea of female promiscuity driving these changes in sperm morphology (Fig. 10.9).[18]

---

[18] It is important to point out that the traits in these analyses were created by binarizing what were initially continuously varying characters. While perhaps not an ideal way to treat these characters, the study still provides a neat example of how the order of trait changes can be inferred using Pagel's method, which is attractive for testing casual, adaptive hypotheses. In cases such as this, the distribution of a continuous character may provide information about whether categorization is justifiable. For example, Holden and Mace (1997) showed that continuous physiological variable lactose digestion capacity (LDC) exhibited a bi-model distribution, therefore making the decision to binarize the trait into high and low LDC populations
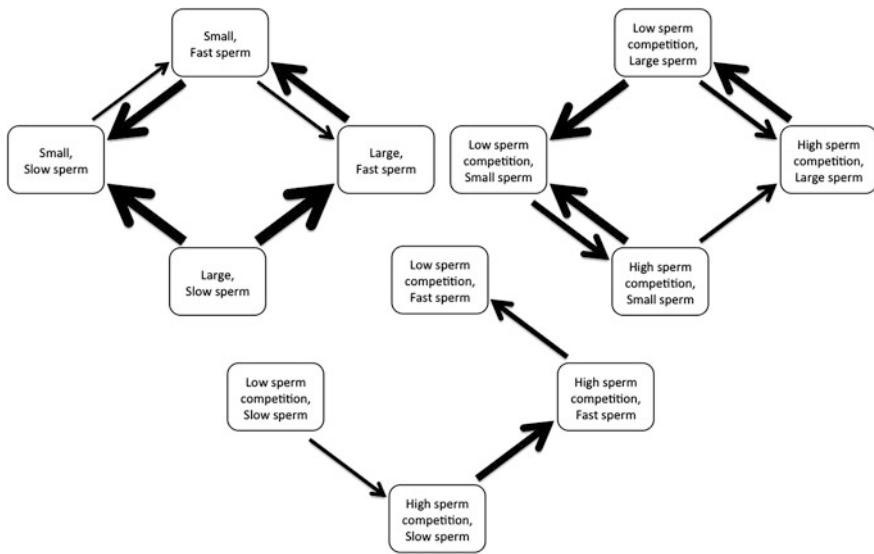
Fig. 10.9 Flow diagrams showing the inferred rates of change from three analyses of sperm competition and sperm characteristics in cichlid fish. Ancestral state reconstructions indicate that the common ancestor of 29 species of these fish had slow, small sperm and experienced low levels of sperm competition. These analyses indicate a clear direction to the order in which the traits change. Sperm gets faster before getting larger (*top left*), and both sperm size (*top right*) and sperm speed (*bottom middle*) increase after sperm competition increases. Figure redrawn from figures in Fitzpatrick et al. (2009) and modified with permission

Other examples of this Bayesian approach have involved studies of mechanisms of sex determination in marine reptiles (Organ et al. 2009), activity periods (Griffin et al. 2012), and social systems in primates (Shultz et al. 2011), and the evolution of brood parasitism in bees (Cardinal et al. 2009). These methods have also been applied to human cultural evolution to examine such things as post-marital residence (Jordan et al. 2009), systems of grammar (Dunn et al. 2011), and political complexity (Currie et al. 2010). It can be seen from the above examples that these model-based techniques are extremely flexible and allow a wide range of evolutionary questions to be addressed. This approach allows extensions to these simple models to be easily incorporated. For example, these basic models assume that the rate of evolution is constant over the tree. However, it is possible to test whether the rate of evolution actually varies over the tree (Penny et al. 2001) or whether the rate of change increases during speciation events (i.e. a punctuated mode of evolution) (Pagel et al. 2006; Pagel 1999).

(Footnote 18 continued)

understandable. Section 10.7 discusses an alternative way to model binary characters that have an underlying continuous distribution.

10.7 Further Issues and Advanced Topics

10.7.1 Stochastic Character Mapping and Alternative Ways to Model Discrete Characters

A further extension of this kind of model-based Bayesian approach is to employ the inferred rate of change to explicitly reconstruct the changes a character undergoes over the whole phylogeny, using a technique known as stochastic character mapping (SCM) (Bollback 2006; Huelsenbeck et al. 2003). We saw earlier how this was traditionally done using parsimony techniques, and character histories can also be approximated with likelihood methods by reconstructing likely states at ancestral nodes and noting where character changes seem to occur between these nodes. SCM goes one step further, by producing a posterior sample of ancestral states and likely changes along the branch, and unlike previous techniques, SCM can show multiple changes along a branch. By explicitly considering the rate of change and branch lengths in these character histories, SCM makes more use of the information available. In addition to this kind of descriptive application, SCM can also be used to detect signatures of positive selection in genetic data and provides an alternative assessment of co-evolution based on the amount of time traits spend together in certain states[19]. Furthermore, the output from SCM analyses can be used in further analyses that examine how other traits are evolving, e.g. variation in rates of change of a continuous character (O'Meara et al. 2006) or the different selective regimes in OU models (Beaulieu et al. 2012).

Here, we have focussed on the Markov chain approach to modelling the evolution of discrete characters. However, it should be noted that other methods are also possible. For example, logistic regression (i.e. a regression analysis in which the dependent variable is a categorical variable) can be adapted to incorporate the variance–covariance structure derived from the phylogenetic associations between species (Ives and Garland 2010; Ives and Garland Chap. 9 this volume). Also, categorical traits can be modelled as if they are related to a continuously varying underlying scale (the *liability*); above a certain value, the trait takes one form, while below this value, it takes another form, etc. (Felsenstein 2005, 2012). One advantage of these methods is that they allow us to examine the covariation between more than just two binary characters.[20] On the other hand, they do not allow us to make direct inferences about the order in which traits change in the manner that Pagel's method does. As with all investigations, the particular method that should be employed will depend on the question being asked and the data available to answer it.

---

[19] SCM is implemented in the program SIMMAP (http://www.simmap.com/) (Bollback 2006).

[20] This is potentially possible in Pagel's method described above, but would be more complicated and involve many more parameters.

10.7.2 Minimum Models and Reversible-Jump MCMC

One practical issue faced by the model-based approach is to find the optimum number of parameters that sufficiently describe the evolutionary process, but do not ask too much of the data. For example, the evolution of a single binary trait may be best described by a simple model in which rates of gains and losses are described by a single parameter, rather than a model in which these two rates are both estimated. So an important task is to try and find minimum models by setting some parameters to zero or making some parameters equal to each other (Pagel and Meade 2005). Even for models with only a few rate parameters, the number of possible ways to do this becomes very large, very quickly. A Bayesian technique known as reversible-jump MCMC (Green 1995; Pagel and Meade 2006) allows us to explore this universe of different models and not only sample parameter values in proportion to their probability, but also sample the different models of evolution themselves, i.e. those models with the optimal number of parameters to explain the data well. By examining the posterior distribution of models, we have a direct way of testing different scenarios of trait evolution, e.g. deciding between dependent and independent models of trait evolution.

10.7.3 Continuous Traits and Packages for Performing Bayesian PCMs

Many other chapters in this book introduce and discuss methods designed for variables that are continuously distributed. The same principles of Bayesian inference can be applied to these methods, too. The main difference is that instead of the rate parameters of a Markov chain, we want to find the parameters of models that involve continuous variables and their evolution. For example, the posterior distributions of the slope and intercept parameters of a linear model (de Villemereuil et al. 2012) or the variance and strength of selection parameters of an OU model (Beaulieu et al. 2012) could be estimated in a Bayesian framework. In addition to employing Markov chain models of discrete trait evolution, the software used in the OPM (BayesTraits) can perform Bayesian implementations of phylogenetic generalized least squares regression and correlation analyses. Additionally, open source software from Bayesian inference using Gibbs sampling (BUGS) and related packages has been used for Bayesian implementations of PCMs, and like BayesTraits, it is able to incorporate phylogenetic uncertainty by performing analyses over a sample of trees (de Villemereuil et al. 2012). Other packages such as MCMCglmm (Hadfield 2010) can also be adapted to handle these kinds of Bayesian analyses in a phylogenetic context (see chap. 11 by de Villemereuil and Nakagawa). Although MCMCglmm cannot currently incorporate more than a single phylogenetic tree into analyses, an attractive feature is that it is flexible enough to deal with traits that can take a number of different distributions.

As a practical consideration, it should be noted that one drawback to these Bayesian methods is that the time required to perform these analyses is greatly increased in comparison with maximum-likelihood methods. This is due in part to the extra time taken in the estimation of the posterior distribution (i.e. the Markov chain may have to run for millions of iterations). The actual computational time will also be greater the more taxa you are analysing or the more parameters you are estimating. This is not only a factor in the final analyses, but also increases the time that needs to be spent in the initial phases of an analysis where suitable priors need to be chosen with care and the inputs that affect how the chain searches parameter space need to be selected. However, the results from a Bayesian analysis are relatively straightforward to handle and interpret, and once familiar with these complications and with enough experience, these issues become less problematic. Processor speeds are increasing ever more rapidly, and the time taken to perform analyses is becoming much more manageable.

10.8 The Practical

In the OPM, you will use the program BayesTraits to perform phylogenetic comparative analyses of discrete characters modelled as a Markov process in a Bayesian MCMC framework. The practical uses the data from Old World monkeys and apes on oestrus advertisement and group composition that you were introduced to above. You will use these data and a sample of phylogenetic trees to test whether these traits have evolved dependently or independently using Pagel's method. At the end of the practical section, we provide a checklist of steps to perform and things to look out for that will help in this process.

10.9 Conclusion

Bayesian techniques for performing phylogenetic comparative analyses provide a powerful and flexible toolkit for tackling a wide range of evolutionary questions. The strengths of this approach lie in its explicit focus on testing between alternative hypotheses,[21] the incorporation of our uncertainty about phylogenetic relationships, and the ability to include prior information to inform our analyses. Furthermore, these Bayesian methods present the results as posterior distributions rather point estimates of likelihoods and parameter values. While this may appear cumbersome, and off-putting for the uninitiated, it is in fact a more realistic representation of what our data can actually tell us about the evolutionary process

---

[21] rather than just focussing on the rejection of null hypotheses as is the case with classical statistical procedures.

that generated them. Becoming familiar with the added complexities that Bayesian analyses entail is well worth the effort as this approach ultimately proves a far more satisfying way of testing evolutionary hypotheses.

Glossary

Bayes factors	Bayes factors are a way of testing between different hypotheses in a Bayesian framework. They are calculated as ratios of the marginal likelihoods of different models. The larger the ratio the more support there is for one model over another. The interpretation of Bayes factors is somewhat arbitrary, but rules of thumb exist to use these values to assess the strength of evidence in favour of one hypothesis over another.
Hyperprior	A hyperprior is a prior distribution on the hyperparameter of a prior distribution. In other words instead of the parameters of the distribution being given fixed values, they themselves are drawn from prior distributions. In the program BayesTraits (which is used in the OPM), the hyperparameters of specified distributions are drawn from uniform distributions. For example, a gamma distribution could have its shape and scale parameters drawn from a uniform distribution ranging from 0 to 10. In comparative analyses, we do not always possess relevant biological information that could inform us about what form and values the priors should take therefore hyperpriors are attractive because they allow us to be less restrictive about the values of a given prior distribution.
Marginal Likelihood	The marginal likelihood of a model is its likelihood scaled by the prior probabilities and integrated over all values of the parameters. In the context of phylogenetic comparative analyses this may also involve integrating over all the trees in the sample.
Markov Chain Monte Carlo (MCMC)	MCMC is a statistical procedure used in Bayesian analyses to search parameter space and sample values in proportion to their posterior probability in order to arrive at an estimate of the posterior distributions of a model and its parameter values. A number of

different criteria can be implemented to govern the way an MCMC searches and samples the posterior distribution. With the *Metropolis–Hasting* algorithm, parameter values that increase the likelihood are always accepted, while those that lead to a decrease are accepted only with a certain probability. The *Gibbs* sampler always accepts proposed values but works by drawing new values from the conditional distributions of the parameters (i.e. the distribution of a parameter given the value of other parameters).

Maximum likelihood

In a maximum-likelihood we search for the values of the parameters of a statistical model that give the largest possible value of the likelihood function.

Prior probability and Priors/Prior distribution

In Bayesian statistics, we need to specify our initial belief about the probability of a hypothesis, given the information available at the time. This belief then gets updated when we gain more information (i.e. this is our belief prior to the assessment of new information). In the context of a comparative analysis, we are assessing the parameters of a statistical model, and before running the analysis and examining the data, we have to specify a prior probability distribution of the values these parameters should take given our current understanding. The chapter by Currie and Meade provides some examples of common prior distributions that are used in comparative analyses. See also *Hyperprior*.

Posterior probability and Posteriors/Posterior distribution

In Bayesian statistics, the posterior probability refers to our belief in a hypothesis after (i.e. posterior to) assessing new information. In the context of a comparative analysis, the results of our analysis give us the posterior probability distribution of values of the parameters of a statistical model. See also *Markov Chain Monte Carlo* (*MCMC*).

References

Arnold C, Matthews LJ, Nunn CL (2010) The 10kTrees website: a new online resource for primate phylogeny. Evol Anthropol 19:114–118

Beaulieu JM, Jhwueng DC, Boettiger C, O'Meara BC (2012) Modeling stabilizing selection: expanding the Ornstein–Uhlenbeck model of adaptive evolution. Evolution 66(8):2369–2383. doi:10.1111/j.1558-5646.2012.01619.x

Bollback JP (2006) SIMMAP: stochastic character mapping of discrete traits on phylogenies. BMC Bioinformatics 7(1):88. doi 10.1186/1471-2105-7-88

Burnham KP, Anderson DR (2002) Model selection and multi-model inference: a practical information-theoretic approach. Springer, New York

Cardinal S, Straka J, Danforth BN (2009) Comprehensive phylogeny of apid bees reveals the evolutionary origins and antiquity of cleptoparasitism. In: Proceedings of the National Academy of Sciences. doi:10.1073/pnas.1006299107

Currie TE, Greenhill SJ, Gray RD, Hasegawa T, Mace R (2010) Rise and fall of political complexity in island south-east Asia and the Pacific. Nature 467(7317):801–804

Davies NB, Krebs JR, West SA (2012) An introduction to behavioural ecology. Wiley, New Jersey

de Villemereuil P, Wells J, Edwards R, Blomberg S (2012) Bayesian models for comparative analysis integrating phylogenetic uncertainty. BMC Evol Biol 12(1):102

Domb LG, Pagel M (2001) Sexual swellings advertise female quality in wild baboons. Nature 410(6825):204–206

Dunn M, Greenhill SJ, Levinson SC, Gray RD (2011) Evolved structure of language shows lineage-specific trends in word-order universals. Nature 473(7345):79–82

Felsenstein J (2005) Using the quantitative genetic threshold model for inferences between and within species. Philos Trans Roy Soc B Biol Sci 360(1459):1427–1434. doi:10.1098/rstb.2005.1669

Felsenstein J (2012) A comparative method for both discrete and continuous characters using the threshold model. Am Nat 179(2):145–156. doi:10.1086/663681

Fitzpatrick JL, Montgomerie R, Desjardins JK, Stiver KA, Kolm N, Balshine S (2009) Female promiscuity promotes the evolution of faster sperm in cichlid fishes. Proc Natl Acad Sci 106(4):1128–1132. doi:10.1073/pnas.0809990106

Green PJ (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82(4):711–732

Griffin RH, Matthews LJ, Nunn CL (2012) Evolutionary disequilibrium and activity period in primates: a bayesian phylogenetic approach. Am J Phys Anthropol 147(3):409–416. doi:10.1002/ajpa.22008

Hadfield JD (2010) MCMC methods for multi-response generalized linear mixed models: the mcmcglmm R package. J Stat Softw 33(2):122

Hibbett DS (2004) Trends in morphological evolution in homobasidiomycetes inferred using maximum likelihood: a comparison of binary and multistate approaches. Syst Biol 53(6):889–903

Holden C, Mace R (1997) Phylogenetic analysis of the evolution of lactose digestion in adults. Hum Biol 69(5):605–628

Huelsenbeck JP, Nielsen R, Bollback JP (2003) Stochastic mapping of morphological characters. Syst Biol 52(2):131–158. doi:10.1080/10635150390192780

Huelsenbeck JP, Rannala B (2003) Detecting correlation between characters in a comparative analysis with uncertain phylogeny. Evolution 57(6):1237–1247

Huelsenbeck JP, Rannala B, Masly JP (2000) Accommodating phylogenetic uncertainty in evolutionary studies. Science 288(5475):2349–2350

Ives AR, Garland T (2010) Phylogenetic logistic regression for binary dependent variables. Syst Biol 59(1):9–26. doi:10.1093/sysbio/syp074

Jordan FM, Gray RD, Greenhill SJ, Mace R (2009) Matrilocal residence is ancestral in Austronesian societies. Proc Roy Soc B Biol Sci 276(1664):1957–1964. doi:10.1098/rspb.2009.0088

Kass RE, Raftery AE (1995) Bayes Factors. J Am Stat Assoc 90(430):773–795

Link WA, Barker RJ (2009) Bayesian inference: with ecological applications. Elsevier Science, New Jersey

Maddison WP (1990) A method for testing the correlated evolution of 2 binary characters—are gains or losses concentrated on certain branches of a phylogenetic tree. Evolution 44(3):539–557

Maddison WP, Maddison DR (2009) Mesquite: a modular system for evolutionary analysis. 2.71 edn

O'Meara BC, Ane C, Sanderson MJ, Wainwright PC (2006) Testing for different rates of continuous trait evolution using likelihood. Evolution 60(5):922–933

Organ CL, Janes DE, Meade A, Pagel M (2009) Genotypic sex determination enabled adaptive radiations of extinct marine reptiles. Nature 461(7262):389–392

Pagel M (1994a) Detecting correlated evolution on phylogenies: a general-method for the comparative-analysis of discrete characters. Proc R Soc Lond Ser B Biol Sci 255(1342):37–45

Pagel M (1994b) Evolution of conspicuous estrous advertisement in old-world monkeys. Anim Behav 47(6):1333–1341

Pagel M (1999) Inferring the historical patterns of biological evolution. Nature 401(6756):877–884

Pagel M, Meade A (2005) Bayesian estimation of correlated evolution across cultures: a case study of marriage systems and wealth transfer at marriage. In: Mace R, Holden CJ, Shennan S (eds) Left Coast Press. Walnut Creek, California

Pagel M, Meade A (2006) Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. Am Nat 167(6):808–825

Pagel M, Meade A, Barker D (2004) Bayesian estimation of ancestral character states on phylogenies. Syst Biol 53(5):673–684

Pagel M, Venditti C, Meade A (2006) Large punctuational contribution of speciation to evolutionary divergence at the molecular level. Science 314(5796):119–121. doi:10.1126/science.1129647

Penny D, McComish BJ, Charleston MA, Hendy MD (2001) Mathematical elegance with biochemical realism: the covarion model of molecular evolution. J Mol Evol 53(6):711–723. doi:10.1007/s002390010258

Posada D (2009) Selecting models of evolution. In: Lemey P, Salemi M, Vandamme AM (eds) The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing. Cambridge University Press, Cambridge, pp 345–361

Ronquist F, van der Mark P, Huelsenbeck JP (2009) Bayesian phylogenetic analysis using MrBayes. In: Lemey P, Salemi M, Vandamme AM (eds) The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing. Cambridge University Press, Cambridge

Sanderson MJ (1993) Reversibility in evolution: a maximum-likelihood approach to character gain loss bias in phylogenies. Evolution 47(1):236–252

Shultz S, Opie C, Atkinson QD (2011) Stepwise evolution of stable sociality in primates. Nature 479(7372):219–222

Chapter 11
General Quantitative Genetic Methods for Comparative Biology

Pierre de Villemereuil and Shinichi Nakagawa

Abstract There is much in common between the aim and tools of the quantitative geneticist and the comparative biologist. One of the most interesting statistical tools of the quantitative genetics (QG) is the mixed model framework, especially the so-called animal model, which can be used for comparative analyses. In this chapter, we describe the phylogenetic generalised linear mixed model (PGLMM), which encompasses phylogenetic (linear) mixed model (PMM). The widely used phylogenetic generalised least square (PGLS) can be seen as a special case of PGLMM. Thus, we demonstrate how PGLMM can be a useful extension of PGLS, hence a useful tool for the comparative biologist. In particular, we show how the PGLMM can tackle issues such as (1) intraspecific variance inference, (2) phylogenetic meta-analysis, (3) non-Gaussian traits analysis, and (4) missing values and data augmentation. Further possible extensions of the PGLMM and applications to phylogenetic comparative (PC) analysis are discussed at the end of the chapter. We provide working examples, using the R package MCMCglmm, in the online practical material (OPM).

11.1 Introduction

Quantitative genetics (QG) and phylogenetic comparative (PC) methods have a lot in common, yet the connections between the two fields have only recently been stressed (Felsenstein 2005; Hadfield and Nakagawa 2010; Stone et al. 2011).

P. de Villemereuil (✉)
Laboratoire d'Écologie Alpine (LECA-UMR CNRS 5553), Université Joseph Fourier,
BP 53, 38041 Grenoble, France
e-mail: bonamy@horus.ens.fr

S. Nakagawa
Department of Zoology, University of Otago, 340 Great King Street,
Dunedin 9054, New Zealand
e-mail: shinichi.nakagawa@otago.ac.nz

Indeed, both frameworks share many characteristics: (1) they aim at the evolutionary study of complex physiological, morphological, or ecological characters for which (2) they assume a Gaussian distribution (but see Sect. 11.3.1 for this assumption to be relaxed), and overall, (3) they aim at compartmentalising the phenotypic variability into an evolutive genetic component and one or several environmentally driven components. Quantitative geneticists have a long history of developing flexible and powerful statistical tools (see Hill and Kirkpatrick 2010, for a historical review), including the so-called 'animal model', which led to statistical developments such as the restricted maximum likelihood (REML) and the framework of (generalised) linear mixed models. Just as comparative phylogeny is using the relationship between species to investigate evolutionary events, the quantitative geneticists are interested into the relationship between individuals to infer the genetic component of polygenic traits. In particular, the 'animal model' is using a pedigree (a comprehensive record of the genealogy of the individuals) to decompose the phenotypic variance into its genetic and environmental components. To do so, the pedigree is transformed into a variance–covariance matrix of relatedness between individuals, which is included as a 'random effect' into the model. We will examine in this chapter how QG tools, namely (generalised) linear mixed models, can be adapted to the PC analysis framework and how they can nicely complement the widely used phylogenetic generalised least square (PGLS; for details of PGLS, see Chaps. 5 and 6). We explain that PGLS can, in fact, be seen as a special case of the phylogenetic (linear) mixed model (PMM) (Lynch 1991), which is, in turn, part of the overarching framework, phylogenetic generalised linear mixed model (PGLMM) (Hadfield and Nakagawa 2010; Ives and Helmus 2011). The evolutionary questions addressed in this chapter will thus be much alike those of the other chapters of Part II.

11.1.1 A Very Brief History of Phylogenetic Mixed Models

Lynch (1991) was the first to recognise the possibility to apply QG methods to comparative analysis, using mixed models to infer phylogeny-wide genetic variances against taxon-specific residual variance. The idea was to replace the variance–covariance matrix of relatedness between individuals by a phylogenetic variance–covariance (or correlation) matrix, which assumes Brownian motion model of trait evolution (i.e. assuming a constant variance in a trait through evolution, so that related species share closer trait values). By doing so, we could estimate ancestral states (or phylogenetic effects) instead of breeding values, and phylogenetic signal (the so-called phylogenetic heritability) instead of pedigree-based heritability Lynch (1991). Despite acknowledged interesting features (e.g. see Miles and Dunham 1993), Lynch's method PMM has only sparsely been highlighted in the PC literature (Housworth et al. 2004; Felsenstein 2008) and it seems to have rarely been used for practical comparative analysis. There are numerous reasons why this is the case. We can, however, come up with two

possible main reasons. First, Felsenstein's (1985) independent contrast (PIC) method had already set a standard for how to analyse inter-species comparative data before Lynch's (1991) QG-based method. Like other biological and human processes, it is likely that a founder-takes-all type of phenomenon has been at work (e.g. Waters et al. 2013). In other words, historical inertia (analogous to phylogenetic inertia!) may have played a role in this neglect on Lynch's important work. Second, unlike PIC, efficient algorithms and easy-to-use implementations have not been available for Lynch's method (at least until recently), even though Housworth et al. (2004) provided some improvement in algorithms, which has especially made estimation for multiresponse (multivariate) models more reliable. The under-usage of PMM feels little ironic because PIC is also a special case of PMM (Housworth et al. 2004).

After two and half decades since Lynch's work, Hadfield and Nakagawa (2010) revived the connections between QG and PC methods by developing a fast computational method for the phylogenetic variance–covariance matrix and its inverse. They have shown how PMM can be implemented in existing R packages (R Development Core Team 2011) and BUGS (Lunn et al. 2000). By developing an MCMC algorithm, they have also extended PMM to PGLMM, which can deal with non-Gaussian characters such as traits following binomial, multinomial, or Poisson distributions. Notably, they have proposed multinomial logit mixed models for a PC method. Such multinomial mixed models have not been used in QG although common in econometrics and political science. They showed that this multinomial PGLMM would be useful for the evolution of multiple discrete traits such as colour polymorphisms (i.e., for example, one taxon having three colour morphs, red, while, and black; see also Sect. 11.2.1). Recently, a number of comparative studies have tackled non-Gaussian traits using the framework of PGLMM (e.g. Ross et al. 2013a, b; Cornwallis et al. 2010; Maklakov et al. 2011).

We believe that it is worthwhile knowing the essence of Hadfield and Nakagawa's algorithm, although it is a little technical, as it represents the key connection between their method and QG animal models. There is a striking similarity between the phylogenetic variance–covariance matrix (hereby noted Σ) and the relatedness matrix (hereby noted $\mathbf{A}$). As stated above, the former represents relatedness among species and is obtained from a phylogenetic tree, whereas the latter represents relatedness among individuals and is obtained from a pedigree. As the matrix $\mathbf{A}$ plays a critical role in estimating additive genetic variance and thus heritability of traits of interest, Σ allows us to estimate the phylogenetic variance and thus phylogenetic signal (see Sect. 11.2). For statistical computation, rather than $\mathbf{A}$ and Σ, we require their inverse matrices, $\mathbf{A}^{-1}$ and Σ^{-1}, whose computation can be extremely slow or even sometimes infeasible (this problem becomes increasingly worse as a pedigree or a phylogeny gets larger). So, the efficient algorithms of animal models (Henderson 1976; Meuwissen and Luo 1992) use the additive genetic variance $\mathbf{S}$, which is an expanded version of $\mathbf{A}$. The matrix $\mathbf{S}$ includes 'missing parents' so that all individuals including ones that do not have parents in an original pedigree will have a set of two parents. Importantly and rather counter-intuitively,

the inverse matrix, $\mathbf{S}^{-1}$, can be computed in much less time than $\mathbf{A}^{-1}$. This inclusion of missing parents is analogous to including the ancestral nodes because a pedigree and a phylogeny share the basic graph structure (with the phylogeny not having fathers). Furthermore, a branch length between parent and child node in a phylogeny is equivalent to inbreeding coefficient represented by a path between two individuals in a pedigree. Therefore, the phylogenetic version of $\mathbf{S}$, say $\mathbf{\Omega}$, can be constructed by including all ancestral nodes (not just tips, i.e. species), and the inverse of this (i.e. $\mathbf{\Omega}^{-1}$) can be used for computation. For example, with a large phylogeny (*ca.* 5,000 species), analysis with $\mathbf{\Sigma}^{-1}$ parametrisation (only using tips) could take over a month while the same analysis with $\mathbf{\Omega}^{-1}$ parametrisation (using tips and nodes) would only be a matter of an hour or so (for more technical details, see Hadfield and Nakagawa 2010).

11.1.2 Roadmap

In this chapter, we will show how QG methods can be useful for (1) multiple measurements data and intraspecific variance inference, (2) phylogenetic meta-analysis framework, (3) PC analysis on non-Gaussian characters, and (4) missing species design, using the framework of missing data theory. The chapter will end with a discussion about the interests and perspective of connections between QG and PC analysis frameworks. Although the sections of this chapter are quite independent from each other, readers who are unfamiliar with mixed models are strongly advised to read the following section. Also, it is recommended reading the following sections in the order of appearance. The reader will find working examples in the online practical material (hereafter OPM) at http://www.mpcm-evolution.com. The two most popular softwares for phylogeny-compatible mixed modelling are the frequentist software ASReml (Gilmour et al. 2006) and the Bayesian R package MCMCglmm (Hadfield 2010). Although the former is much faster, the OPM focus on the second package for several reasons. To begin with, MCMCglmm being Bayesian, it is more flexible than its frequentist equivalent and, in particular, it has better properties regarding non-Gaussian traits (de Villemereuil et al. 2013). Perhaps most importantly, the syntax of MCMCglmm is more oriented towards PC analysis and Hadfield and Nakagawa's (2010) algorithm has been directly implemented in it.

11.2 First Step: Mixed Model for Multiple Measurement Data

Random effects are commonly used within the mixed models framework to account for non-independent structure in the 'residuals'. In the context of comparative analysis, it can be useful to use such random effects to take phylogenetic

relationship between species into account. This section will constitute an intro-
duction to mixed models and their applications to comparative analysis, by using
the common case of multiple measurement data and intraspecific variance infer-
ence. For theoretical developments and review of methods for intraspecific vari-
ability, please refer to Chap. 7.

11.2.1 Description of the Simple Model

Let us assume we have phenotypic data $\mathbf{y}$ (e.g. body size) for several species and
co-factors of interest (we will assume just one called $\mathbf{x}$, e.g. the temperature of the
environment). Now, consider we also have a phylogeny from which we derived a
phylogenetic correlation matrix $\mathbf{\Sigma}$ (say using the classical Brownian motion
assumption[1]). How can we define a mixed model to infer a relationship between
$\mathbf{y}$ and $\mathbf{x}$ while taking the phylogenetic structure into account? The model would be
as follows:

$$\mathbf{y} = \mu + \beta\mathbf{x} + \mathbf{a} + \mathbf{e} \tag{11.1}$$

where μ and β, respectively, are the intercept and the slope for the co-factor[2] $\mathbf{x}$, $\mathbf{a}$ is
the phylogenetic random effect, and $\mathbf{e}$ is the residual error. Now, the two last terms
are assumed to be normally distributed with:

$$\begin{aligned} \mathbf{a} &\sim \mathcal{N}\left(0, \sigma_P^2 \mathbf{\Sigma}\right) \\ \mathbf{e} &\sim \mathcal{N}(0, \sigma_R^2 \mathbf{I}) \end{aligned} \tag{11.2}$$

where $\mathbf{I}$ stands for the relevant identity matrix. Our model, thus, assumes that
phylogenetic effects are correlated according to the phylogenetic correlation
matrix $\mathbf{\Sigma}$. Note also that our model is estimating two variances: V_P is the variance
of the phylogenetic effect and V_R is the residual error (environment effects,
intraspecific variance, measurement error, etc.).

It is important here to stress the resemblances and dissimilarities between the
PGLMM above, and the classical model assumed in PGLS is denoted as:

$$\mathbf{y} \sim \mathcal{N}\left(\mu + \beta\mathbf{x}, \sigma_P^2 \mathbf{\Sigma}\right) \tag{11.3}$$

---

[1] But, any kind of evolutionary model yielding such a variance–covariance matrix can be used,
such as Martins and Hansen's (1997) or ACDC processes (Blomberg et al. 2003). In practice,
parameters of such models would be inferred before using the mixed model, but nothing, in
theory, forbids the construction of a complex mixed model inferring these components along
with performing the comparative regression.

[2] Of course, there can be an arbitrary number of such co-factors (either continuous or categorical
variables).

Although the models are very much alike, a striking difference is the absence of the residual term e in the PGLS model, which only estimates σ_P^2, but not σ_R^2. In PC analysis, this constraint (that the residuals are distributed exactly according to the phylogeny) is usually relaxed using phylogenetic signal inference and introducing an extra parameter whose role is to measure such signal. By contrast, quantitative geneticists always assume that the pedigree (hence, the genetics) is only one source of the observed variability, the other one being the environment, usually captured by the residuals. Fortunately, comparative biologists do not have to give up on their usual tools to consider using mixed models. The model described in Eqs. 11.1 and 11.2 is equivalent to Pagel's λ model of phylogenetic signal inference (Freckleton et al. 2002; Housworth et al. 2004), given that the matrix Σ is a correlation matrix (i.e. diagonal elements are equal to 1, Hansen and Orzack 2005; Hadfield and Nakagawa 2010). Indeed, very much alike the heritability for QG analysis, we can define Lynch's phylogenetic heritability $\lambda = \frac{\sigma_P^2}{\sigma_P^2 + \sigma_R^2}$ as a measure of the phylogenetic signal.[3] Actually, the above difference between PGLMM and PGLS is the only major one. Most other differences actually lie on which extensions of this model are used. For example, random effects and hierarchical modelling for non-Gaussian traits (see Sect. 11.3.1) are widely used in the field of QG, but scarcely in PC analyses. This chapter, among other things, aims at demonstrating how some of the quantitative geneticist 'tools' can prove to be useful to the comparative biologist.

11.2.2 Using Random Effects: The Case of Multiple Measurements

In many comparative cases, we have multiple measurements for each species. An extension to deal with such cases is straightforward, and we have:

$$\mathbf{y} = \mu + \beta \mathbf{x} + \mathbf{a} + \mathbf{s} + \mathbf{e} \tag{11.4}$$

$$\mathbf{s} \sim \mathcal{N}(0, \sigma_S^2 \mathbf{I}) \tag{11.5}$$

where $\mathbf{s}$ is the 'multiple measurement effect' or species-specific effect after taking out the phylogenetic effect. This effect accounts for the variability that has been

---

[3] Note that, although λ could be forced to one by setting up $\sigma_R^2 = 0$ in the model, this could cause numerical instability in frequentist software or strong auto-correlation in MCMC algorithms. The software MCMCglmm, for example, does not allow such a setting. Furthermore, there is some relevance in assuming that some of the biological variability is not captured by the phylogeny (such as environment or even measurement variability), hence assuming a residual variance. Also, notably, when $\sigma_R^2 = 0$, PMM can be seen as equivalent to PGLS and thus PIC (Stone et al. 2011; Blomberg et al. 2012).

caused by the species' contingent characteristics (or species-specific effects). σ_S^2 is the variance of this effect. The other symbols are as in Eqs. 11.1 and 11.2.

Together, σ_P^2 and σ_S^2 accounts for the between-species variability (the first being caused by the evolutionary history, the second by contingent events). By contrast, the residual term σ_R^2 is a measure of the intraspecific variance of the trait.[4] Note that we are assuming the same intraspecific variance for all the species in the dataset, which might be considered as a very strong (although practical) assumption.

A careful inspection of Eq. 11.4 might reveal a troubling fact. As it stands, we have no clue of the type of relationship the slope β is measuring. As a comparative biologist, the reader would most likely be interested in the between-species slope. If the co-factor $\mathbf{x}$ only contains only one value per species (or mean specific values), then there is no problem, since for an individual j belonging to species i, the Eq. 11.4 can be rewritten as follows:

$$y_{i,j} - a_i - s_i = \mu + \beta x_i + e_{i,j} \tag{11.6}$$

Hence, we can consider the random effect a_i and s_i as within-species centring effects and the slope β as a between-species slope.

Things are slightly more complicated using individual measurements in $\mathbf{x}$, but it is still possible to obtain the between-species and within-species slopes using a technique called *within-group centring* (Davis et al. 1961; van de Pol and Wright 2009). The principle of this technique is to separate the predictor $\mathbf{x}$ into two components: one containing the group-level mean of $\mathbf{x}$ (here, the specific mean) and a second one containing the within-group variability. For an individual j belonging to species i, the new model would thus be:

$$y_{i,j} = \mu + \beta_B \bar{x}_i + \beta_W (x_{i,j} - \bar{x}_i) + a_i + s_i + e_{i,j} \tag{11.7}$$

where:

$$\bar{x}_i = \frac{1}{J_i} \sum_{j=1}^{J_i} x_{i,j} \tag{11.8}$$

for J_i being the number of individuals in species i. Here, we are thus fitting two slopes: β_B is the slope of regression between species and β_W is the (common) slope of regression within each species. The model could be further complicated to include one slope per species (a so-called *random slope model*), but such a complex model would be out of the scope of this chapter. Note that, by construction, the predictors $\bar{x}_i$ and $(x_{i,j} - \bar{x}_i)$ are perfectly orthogonal. Therefore, β_B and β_W are truly independent. Finally, the calculation of λ would be changed to account for the extra random effect:

---

[4] This is not totally true, since σ_R^2 also include noise such as measurement error, which is very difficult to distinguish from intraspecific variance without a careful design.

$$\lambda = \frac{\sigma_P^2}{\sigma_P^2 + \sigma_S^2 + \sigma_R^2} \qquad (11.9)$$

We are thus able to estimate intraspecific variance and between-species slope for multiple measurement data with the help of a new random effect. This is only a particular demonstration of the utility of multiple random effects model. One could also use them to account for problems of unbalanced sampling in one's dataset: spatial correlation, biogeographic regions, etc. (see Ives and Zhu 2006). In theory, most of the dependency structures in the error of the model could be accounted for by a random effect.

11.2.3 Phylogenetic Meta-Analysis Using Random Effects

Meta-analysis is a powerful statistical tool to combine weighted results of multiple studies on the same or similar topics. As such, although the technique originated from medical and social sciences, meta-analysis has been used extensively in the field of ecology and evolution (Nakagawa and Poulin 2012; Koricheva et al. 2013). In ecological or evolutionary meta-analysis, it is common that data include multiple species, and therefore, the data look similar to those of comparative analysis. The main difference is that what are 'traits' in PC analysis (e.g. brain size) are 'effect sizes' in meta-analysis (e.g. a relationship between brain size and reproductive success within a species). Such effect sizes are commonly standardised statistical metrics, which are dimensionless (Cohen 1988; Nakagawa and Cuthill 2007), so that they can be compared across studies or species. Four commonly used effect size metrics[5] are: (1) Fisher's z-transformation of correlation coefficient (Zr), (2) Hedges' d and its variants, (3) response ratio on the natural logarithm ($\ln R$), and (4) odds ratio on the natural logarithm ($\ln OR$) (Nakagawa and Santos 2012; Koricheva et al. 2013). A recent study suggests the importance of incorporating phylogeny in meta-analysis because meta-analytic models with and without phylogeny could result in different conclusions (Chamberlain et al. 2012). Here, we will describe phylogenetic meta-analytic models. A working example of such analysis can be found in the OPM.

Several versions of phylogenetic meta-analysis have been proposed (Adams 2008; Lajeunesse 2009; Hadfield and Nakagawa 2010). Although they are slightly different in their details, they all aim for incorporating phylogenetic non-independence. Here, we describe the one based on PMM, described in Hadfield and Nakagawa (2010). In a phylogenetic meta-analytic, we have a vector of effect sizes $\mathbf{z}$ and each effect size has its sampling error variance (all stored in a vector $\mathbf{v_m}$),

---

[5] These standardised metrics are unbounded and follow approximately normal distributions. However, note that the correlation coefficient r is bounded at -1 and 1 and does not follow a normal distribution.

which may sometimes be referred to as measurement error variance. The model is denoted as:

$$\mathbf{z} = \mu + \mathbf{a} + \mathbf{m} + \mathbf{e} \tag{11.10}$$

$$\mathbf{m} \sim \mathcal{N}(0, \mathbf{v_m} \mathbf{I}) \tag{11.11}$$

where μ is the meta-analytic (grand) mean, $\mathbf{m}$ are the sampling (measurement) error effects,[6] and all the other symbols are as in Eqs. 11.1 and 11.2. Sampling error variance is assumed to be known, and for all common effect size statistics, equations are available to obtain their sampling error variances. For example, the sample variance for Zr is $\frac{1}{n-3}$, where n is sample size used to estimate a correlation coefficient. Also note that meta-analysis is typically an intercept model (i.e. μ is the only fixed factor estimated). This is because the main purpose of a meta-analysis is to identify a general trend. But, as you may realise, it is easy to add a predictor (co-factor $\mathbf{x}$ in Eq. 11.1), by building up on Eq. 11.10. This model is expressed as:

$$\mathbf{z} = \mu + \beta \mathbf{x} + \mathbf{a} + \mathbf{m} + \mathbf{e}. \tag{11.12}$$

This model is known as (phylogenetic) meta-regression, and we can add as many fixed factors and random factors as required. Such mixed model-based phylogenetic meta-analysis or meta-regression has recently been used in a number of studies (e.g. Horváthová et al. 2012; Santos and Nakagawa 2012; Prokop et al. 2012; Garamszegi et al. 2012).

Incidentally, remember the example in Sect. 11.2. There, we had multiple measurements per species. Rather than all these raw multiple measurements, let us suppose that we only have species trait means and standard errors (or alternatively, standard deviations and sample sizes). In such a case, we have a model where the square of standard errors for each species trait value can be considered as sampling error variances $\mathbf{v_m}$ so that:

$$\mathbf{y} = \mu + \beta \mathbf{x} + \mathbf{a} + \mathbf{m} + \mathbf{e} \tag{11.13}$$

where $\mathbf{y}$ is a vector of the trait mean for each species. Now, we can see that a phylogenetic meta-regression (Eq. 11.12) and a PC analysis incorporating within-species (sampling error) variance (Eq. 11.13) are mathematically equivalent (see Chap. 7); such equivalence has been pointed out previously (Nakagawa and Santos 2012; Jennions et al. 2012).

---

[6] In a typical non-phylogenetic meta-analysis, a unit of analysis is 'study' where one effect size is taken from one study. Here, we assume that one effect size from each species comes from one study or $n_{\text{effect}} = n_{\text{species}} = n_{\text{study}}$.

11.3 Extensions of the Phylogenetic Linear Mixed Model

11.3.1 Non-Gaussian Characters: Generalised Linear Mixed Model

One of the main advantages of the mixed model framework is that it has been generalised to non-Gaussian response distribution (Gilmour et al. 1985; Breslow and Clayton 1993). As a result, it is now relatively easy to investigate non-Gaussian comparative data using a generalised phylogenetic mixed model (although one needs to be aware of the classical pitfalls, see Bolker et al. 2009). Because of the flexibility of the MCMC algorithm, the MCMCglmm package is one of the most comprehensive in terms of the available distributions (even some complex ones like zero-inflated Poisson). Note that non-Gaussian traits can also be analysed using ASReml software with common distributions such as Poisson and binomial.[7]

Mixed models are generalised by adding a hierarchical layer in the model described in Eq. 11.1. Indeed, we begin by assuming a hypothetical latent trait $\mathbf{l}$, which satisfies Eq. 11.1:

$$\mathbf{l} = \mu + \beta\mathbf{x} + \mathbf{a} + \mathbf{e} \tag{11.14}$$

Note that all assumptions detailed in Sect. 11.2 hold for $\mathbf{l}$ here, since we are using the same model.[8] Then, we use a 'link function' g to draw the relationship between this latent trait $\mathbf{l}$ and our actual non-Gaussian data $\mathbf{y}$. This function will transform $\mathbf{l}$ into a quantity $g^{-1}(\mathbf{l})$, which will be the expectation of the distribution of $\mathbf{y}$.

For example, for count data, we will assume a Poisson distribution (noted $\mathscr{P}$), which only accepts positive mean. The canonical link function is the logarithm, so that:

$$\mathbf{y} \sim \mathscr{P}(\exp(\mathbf{l})) \tag{11.15}$$

You will find a working example in the OPM using count data and Poisson distribution.

For dichotomous (binary) data, we will assume a binomial distribution (noted $\mathscr{B}$) with a vector of probability of "success", $\mathbf{p}$:

$$\mathbf{y} \sim \mathscr{B}(\mathbf{p}) \tag{11.16}$$

---

[7] However, the penalised quasi-likelihood used in ASReml has been shown to largely underestimate the variance components for binary traits (Gilmour et al. 2006; de Villemereuil et al. 2013).

[8] Note, however, that $\mathbf{e}$ in Eq. 11.14 should be considered as the effect due to additive dispersion rather than the residuals (for additive dispersion, see Nakagawa and Schielzeth 2010).

One of the two link functions is usually assumed: the logit or the probit functions. The logit link function is the canonical link function, calculated as:

$$\mathbf{l} = \log(\frac{\mathbf{p}}{1 - \mathbf{p}}) \tag{11.17}$$

The probit link function is actually the inverse function of the cumulative distribution function of a standard normal distribution, noted Φ:

$$\mathbf{l} = \Phi^{-1}(\mathbf{p}) \tag{11.18}$$

The probit has less mathematical support than the logit, but it has a better biological interpretation, due to its strong link with the long-standing *threshold model* (Wright 1934; Dempster and Lerner 1950), widely used in QG for dichotomous traits (see also Chap. 9 for the use of phylogenetic logistic regression for binary traits).

Because of this hierarchical modelling, the generalisation of the phylogenetic mixed model to non-Gaussian traits allows us to keep assumptions about the main Gaussian evolution processes, for example the Brownian motion model. Regarding dichotomous traits, this model is indeed philosophically different from the Markovian processes widely used in the field of PC analysis (see Chaps. 10 and 16). Whereas the latter model discretise the genetics of the dichotomous trait in assuming probabilities of 'jump' from one state to the other, the former assume a polygenic basis of the trait and thus a 'smoother' evolution (and possibly intraspecific variability). More generally, a Gaussian latent trait is justified if you consider highly polygenic characters (even for the discrete ones!), because then the result of all genetic effects should be normally distributed, according to Fisher's (1918) infinitesimal model. Another interest lies, of course, in the fact that most kinds of distributions can be used in Eq. 11.15. Indeed, most ecologically interesting distributions are available in the MCMCglmm package (binomial, ordinal, multinomial, Poisson, zero-inflated Poisson, etc.).

11.3.2 Missing Species and Data Augmentation

When we compile comparative data of a certain taxon, it is difficult to get all trait values for every species in that taxon. In other words, missing data are commonplace in comparative data. What most of researchers have been doing is to ignore species in which trait information is missing and to conduct analysis without such species—this is called 'complete-case analysis'. Unfortunately, complete-case analysis in comparative data will often result in biased estimates (Nakagawa and Freckleton 2008; Garamszegi and Møller 2011). To understand why this is so, we will benefit from learning some basics on missing data mechanisms.

Missing data mechanisms in the statistical literature are merely a classification of how missing data are related to observed data so that missing data mechanisms do not imply casual explanations for missing data. A series of work by Rubin and Little has set foundations for missing data theory (Rubin 1976, 1987; Little and Rubin 2002). In missing data theory, three types of missing data are recognised: (1) missing completely at random (MCAR), (2) missing at random (MAR), and (3) missing not at random (MNAR). To understand these mechanisms, we need to introduce three concepts, by using notations by Enders (2010). First, we need to recognise that the data matrix $\mathbf{Y}$ can be decomposed into an observed part $\mathbf{Y}_o$ and a missing part $\mathbf{Y}_m$. Second, the matrix $\mathbf{R}$ that has the same dimension as $\mathbf{Y}$ has either 0 or 1 in its elements, with 0 signifying 'observed' and 1 'missing'. The matrix $\mathbf{R}$ is referred to as missingness. Third, we call $\boldsymbol{\theta}$ the vector of parameters that described the relationships between the data $\mathbf{Y}$ and its missingness $\mathbf{R}$. Now, the probability distribution function for MCAR can be written as:

$$p(\mathbf{R}|\boldsymbol{\theta}) \tag{11.19}$$

It reads as follows: the probability of whether an element in $\mathbf{R}$ takes 0 or 1 depends neither on $\mathbf{Y}_o$ nor on $\mathbf{Y}_m$. The lack of links between $\mathbf{R}$ and the data $\mathbf{Y}$ is solely described by $\boldsymbol{\theta}$ (cf. Enders 2010). That is, missing values in a variable of interest are distributed completely at random in relation to any other variables. This function also implies that when missing data is MCAR, the complete-case analysis will provide unbiased results although statistical power may be reduced. However, MCAR is a very strong and actually unrealistic assumption because biological processes usually cause missing values. For example, we are less likely to have life-history data on rare species than on more abundant ones. Therefore, a more realistic assumption for missing data is MAR. The probability distribution function for MAR is:

$$p(\mathbf{R}|\mathbf{Y}_o, \boldsymbol{\theta}) \tag{11.20}$$

Again, it reads as follow: the probability of having 0 or 1 in $\mathbf{R}$ depends only on $\mathbf{Y}_o$, and this relationship between $\mathbf{R}$ and $\mathbf{Y}$ is described by $\boldsymbol{\theta}$. That is, missing values in a variable of interest are due to another variable that we have complete data on. For example, missing values in life-history data of rarer species are MAR, if we have complete abundance data, which governs the probability of missingness. It is notable that MAR does not really mean 'MAR' in its usual sense (although this might be very confusing).

Similarly, the probability distribution function for MNAR is:

$$p(\mathbf{R}|\mathbf{Y}_o, \mathbf{Y}_m, \boldsymbol{\theta}) \tag{11.21}$$

It reads as follow: the probability of taking 0 or 1 in $\mathbf{R}$ depends both on $\mathbf{Y}_o$ and $\mathbf{Y}_m$, and this relationship between $\mathbf{R}$ and $\mathbf{Y}$ is described by $\boldsymbol{\theta}$. That is, missing values in a variable of interest are due to another variable that we do not have any

information on or missing values are due to missing values themselves. For example, missing data in life-history data of rarer species are MNAR, if we do not have abundance data, which governs missingness, or if a life-history trait makes particular species rare so that it is difficult to obtain such life-history data of rare species. Importantly, if we conduct complete-case analysis on MAR or MNAR missing data, parameter estimates will often be biased. Therefore, all comparative analyses not accounting for missing species could provide biased parameters and potentially lead to incorrect conclusions, because it is unlikely that such missing species comply with the MCAR assumption. A working example using MCMCglmm that assumes MAR can be found in the OPM.

An important problem here is that we never really know how MNAR missing values were created because the patterns of missingness depend on missing values themselves. Thus, treating MNAR missingness is usually very difficult or often infeasible. Therefore, the most practical assumption is MAR, and many methods to deal with missing data under MAR have been developed (Enders 2010; van Buuren 2012). One notable method is data augmentation using Bayesian MCMC. Although methodological details are beyond the scope of this chapter (for an accessible account, see Enders 2010), data augmentation will provide unbiased parameter estimates when missing data in a dataset fulfils MAR. MCMCglmm uses a data augmentation method when missing data are in the response variable, but not predictors. Datasets with missing data in multiple variables need to resort to either using other methods such as multiple imputation (van Buuren 2012) or using multiresponse models, which ASReml and MCMCglmm are capable of running (for details, see Gilmour et al. 1985; Hadfield 2010). Although there are still a handful of examples of comparative studies utilising the missing data theory (e.g. Fisher et al. 2003; Gonzalez-Suarez et al. 2012; Cleasby and Nakagawa 2012), we expect that the use of missing data missing data augmentation or other methods for dealing with missing data such as multiple imputation will be commonplace in PC analysis in the near future.

11.4 Discussion

Throughout this chapter, we have seen some of the most interesting properties of PGLMM, arising from using QG mixed models for PC analysis. Because mixed models are nowadays a standard in ecology and evolution, and the (sometimes quite philosophical) difference between fixed and random effects is relatively well understood by the scientific community (reviewed in Gelman and Hill 2006), we expect that a shift towards phylogenetic mixed model will not be a huge step for any practitioner interested in evolution. Such a shift would not represent a significant change in the underlying model, since there is a strong proximity between phylogenetic mixed models and phylogenetic generalised least squares. Yet, the

GLMM framework might allow one to tackle some of the important issues listed in the sections above. Incidentally, note that all the examples in this chapter use one response variable, but all of the models above can be easily extended to be multiresponse models, for which MCMCglmm provides an implementation (Hadfield and Nakagawa 2010; Nakagawa and Santos 2012), as well as ASReml (Gilmour et al. 2006). Additionally, multiresponse models for non-Gaussian traits can account for a different distribution for each trait (Hadfield 2010).

Regarding the implementation of the models, we focused on the MCMCglmm package in this chapter. Yet, although the Bayesian MCMC algorithm possesses numerous advantageous sides in the case of GLMM implementation, it is important to stress that one does not need to be Bayesian to run such models.[9] However, we are still lacking a nice and flexible frequentist package that can readily be used in the context of PGLMM (but see Ives and Helmus 2011), mainly because accurate likelihood-based estimations for generalised linear mixed models have proven difficult and still are an active area of research in statistics.

Further extensions of the phylogenetic mixed model are to be expected in the future. For example, Stone et al. (2011) suggested, among other methods, the use of MCMCglmm to control for population structure within species. On the same idea, Buckley et al. (2010) used MCMCglmm to fit 'species within the phylogeny', 'populations within species', and 'years within population' effects. Such nested design is easy and quite intuitive to implement using random effects in a mixed model context (Schielzeth and Nakagawa 2013). Also, one can incorporate two different phylogenies as random effects to investigate traits whose evolution is affected by interactions between species, such as host–parasite, plant–pollinator, and predator–prey interactions (Rafferty and Ives 2013; Hadfield et al. 2014). Another way to extend the mixed model framework would be to allow for uncertainty in the relatedness matrix. In the context of PGLS, de Villemereuil et al. (2012) showed that failing to take this uncertainty into account could lead to anti-conservative standard error of estimates. Although technically difficult to implement efficiently in a package such as MCMCglmm (J. Hadfield, personal communication, but see Ross et al. 2013a, about an implementation of such uncertainty using MCMCglmm), sampling in a distribution of phylogenies instead of using the consensus, one might lead to better estimations (Huelsenbeck and Rannala 2003; de Villemereuil et al. 2012). Because pedigrees are also not known without error (either being social or genetically determined, see Charmantier and Réale 2005; Sillanpää 2011), such 'phylogenetic/animal models with uncertainty' would benefit both quantitative genetic and comparative analysis statistical fields.

---

[9] The data augmentation of Sect. 11.3.2, though, is very much linked to the MCMC algorithm.

References

Adams D (2008) Phylogenetic meta-analysis. Evolution. Int J Organ Evol 62(3):567–572. doi:10.1111/j.1558-5646.2007.00314.x

Blomberg SP, Garland T, Ives AR (2003) Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution 57(4):717–745

Blomberg SP, Lefevre JG, Wells JA, Waterhouse M (2012) Independent contrasts and PGLS regression estimators are equivalent. Syst Biol 61(3):382–391. doi:10.1093/Sysbio/Syr118

Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, White JSS (2009) Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol Evol 24(3):127–135. doi:10.1016/j.tree.2008.10.008

Breslow NE, Clayton DG (1993) Approximate inference in generalized linear mixed models. J Am Stat Assoc 88(421):9–25. doi:10.2307/2290687

Buckley YM, Ramula S, Blomberg SP, Burns JH, Crone EE, Ehrlén J, Knight TM, Pichancourt JB, Quested H, Wardle GM (2010) Causes and consequences of variation in plant population growth rate: a synthesis of matrix population models in a phylogenetic context. Ecol Lett 13(9):1182–1197. doi:10.1111/j.1461-0248.2010.01506.x

Chamberlain S, Hovick S, Dibble C, Rasmussen N, Van Allen B, Maitner B, Ahern J, Lukas B, Roy C, Maria M, Carrillo J, Siemann E, Lajeunesse M, Whitney K (2012) Does phylogeny matter? Assessing the impact of phylogenetic information in ecological meta-analysis. Ecol Lett 15(6):627–636. doi:10.1111/j.1461-0248.2012.01776.x

Charmantier A, Réale D (2005) How do misassigned paternities affect the estimation of heritability in the wild? Mol Ecol 14(9):2839–2850. doi:10.1111/j.1365-294X.2005.02619.x

Cleasby IR, Nakagawa S (2012) The influence of male age on within-pair and extra-pair paternity in passerines. Ibis 154(2):318–324. doi:10.1111/J.1474-919x.2011.01209.X

Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd edn. Lawrence Erlbaum, Hillsdale, New Jersey

Cornwallis CK, West SA, Davis KE, Griffin AS (2010) Promiscuity and the evolutionary transition to complex societies. Nature 466(7309):969–72. doi:10.1038/nature09335

de Villemereuil P, Wells JA, Edwards RD, Blomberg SP (2012) Bayesian models for comparative analysis integrating phylogenetic uncertainty. BMC Evol Biol 12(1):102. doi:10.1186/1471-2148-12-102

de Villemereuil P, Gimenez O, Doligez B (2013) Comparing parent–offspring regression with frequentist and bayesian animal models to estimate heritability in wild populations: a simulation study for gaussian and binary traits. Meth Ecol Evol 4(3):260–275. doi:10.1111/2041-210X.12011

Davis J, Spaeth J, Huson C (1961) A technique for analyzing the effects of group composition. Am Sociol Rev 26(2):215–225. doi:10.2307/2089857

Dempster ER, Lerner IM (1950) Heritability of threshold characters. Genetics 35(2):212–236

Enders CK (2010) Applied missing data analysis. Methodology in the social sciences. Guilford Press, New York, 2010008465 GBB060973 Craig K. Enders. ill.; 26 cm. Includes bibliographical references (p 347–358) and indexes. Methodology in the social sciences

Felsenstein J (1985) Phylogenies and the comparative method. Am Nat 1–15

Felsenstein J (2005) Using the quantitative genetic threshold model for inferences between and within species. Philos Trans: Biol Sci 360(1459):1427–1434

Felsenstein J (2008) Comparative methods with sampling error and within-species variation: contrasts revisited and revised. Am Nat 171(6):713–725. doi 10.1086/587525

Fisher D, Blomberg S, Owens I (2003) Extrinsic versus intrinsic factors in the decline and extinction of Australian marsupials. Proc Biol Sci/Roy Soc 270(1526):1801–1808. doi:10.1098/rspb.2003.2447

Fisher RA (1918) The correlation between relatives on the supposition of Mendelian inheritance. Trans Roy Soc Edinb 52:399–433

Freckleton RP, Harvey PH, Pagel M (2002) Phylogenetic analysis and comparative data: A test and review of evidence. Am Nat 160(6):712–726. doi:10.1086/343873

Garamszegi LZ, Møller AP (2011) Nonrandom variation in within-species sample size and missing data in phylogenetic comparative studies. Syst Biol 60(6):876–880

Garamszegi LZ, Marko G, Herczeg G (2012) A meta-analysis of correlated behaviours with implications for behavioural syndromes: mean effect size, publication bias, phylogenetic effects and the role of mediator variables. Evol Ecol 26(5):1213–1235. doi:10.1007/S10682-012-9589-8

Gelman A, Hill J (2006) Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, Cambridge

Gilmour AR, Anderson RD, Rae AL (1985) The analysis of binomial data by a generalized linear mixed model. Biometrika 72(3):593–599. doi:10.1093/biomet/72.3.593

Gilmour AR, Gogel BJ, Cullis BR, Thompson R (2006) ASReml user guide release 2.0. http://www.vsni.co.uk/software/asreml/

Gonzalez-Suarez M, Lucas PM, Revilla E (2012) Biases in comparative analyses of extinction risk: mind the gap. J Anim Ecol 81(6):1211–1222

Hadfield JD (2010) MCMC methods for multi-response generalised linear mixed models: The MCMCglmm R package. J Stat Softw 33(2):1–22

Hadfield JD, Nakagawa S (2010) General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. J Evol Biol 23(3):494–508. doi:10.1111/j.1420-9101.2009.01915.x

Hadfield JD, Kranov B, Poulin R, Nakagawa S (2014) A tale of two phylogenies: comparative analyses of ecological interactions. Am Nat 183(2):174–187

Hansen TF, Orzack SH (2005) Assessing current adaptation and phylogenetic inertia as explanations of trait evolution: the need for controlled comparisons. Evolution 59(10):2063–2072

Henderson C (1976) A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics 1:69–83

Hill WG, Kirkpatrick M(2010) What animal breeding has taught us about evolution. Ann Rev Ecol, Evol Syst 41:1–19. doi:10.1146/annurev-ecolsys-102209-144728

Horváthová T, Nakagawa S, Uller T (2012) Strategic female reproductive investment in response to male attractiveness in birds. Proc Roy Soc B-Biol Sci 279(1726):163–170

Housworth E, Martins E, Lynch M (2004) The phylogenetic mixed model. Am Nat 163(1):84–96. doi:10.1086/380570

Huelsenbeck JP, Rannala B (2003) Detecting correlation between characters in a comparative analysis with uncertain phylogeny. Evolution 57(6):1237–1247

Ives AR, Helmus MR (2011) Generalized linear mixed models for phylogenetic analyses of community structure. Ecol Monogr 81(3):511–525

Ives AR, Zhu J (2006) Statistics for correlated data: phylogenies, space, and time. Ecol Appl 16(1):20–32

Jennions MD, Kahn AT, Kelly CD, Kokko H (2012) Meta-analysis and sexual selection: past studies and future possibilities. Evol Ecol 26(5):1119–1151. doi: 10.1007/S10682-012-9567-1

Koricheva J, Gurevitch J, Mengersen K (2013) The handbook of meta-analysis in ecology and evolution. Princeton University Press, Princeton

Lajeunesse M (2009) Meta-analysis and the comparative phylogenetic method. The Am Nat 174(3):369–381. doi:10.1086/603628

Little RJA, Rubin DB (2002) Statistical analysis with missing data, Wiley series in probability and statistics, 2nd edn. Wiley, Hoboken, N.J., p 349–364

Lunn DJ, Thomas A, Best N, Spiegelhalter D (2000) Winbugs—a bayesian modelling framework: concepts, structure, and extensibility. Stat Comput 10(4):325–337

Lynch M (1991) Methods for the analysis of comparative data in evolutionary biology. Evolution 45(5):1065–1080. doi:10.2307/2409716

Maklakov AA, Immler S, Gonzalez-Voyer A, Ronn J, Kolm N (2011) Brains and the city: big-brained passerine birds succeed in urban environments. Biol Lett 7(5):730–732. doi:10.1098/Rsbl.2011.0341

Martins EP, Hansen TF (1997) Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. Am Nat 149(4):646–667

Meuwissen T, Luo Z (1992) Computing inbreeding coefficients in large populations. Genet Sel Evol 24:305–313. doi:10.1186/1297-9686-24-4-305

Miles DB, Dunham AE (1993) Historical perspectives in ecology and evolutionary biology: the use of phylogenetic comparative analyses. Ann Rev Ecol Syst 587–619

Nakagawa S, Cuthill IC (2007) Effect size, confidence interval and statistical significance: a practical guide for biologists. Biol Rev 82(4):591–605

Nakagawa S, Freckleton RP (2008) Missing inaction: the dangers of ignoring missing data. Trends Ecol Evol 23(11):592–596

Nakagawa S, Poulin R (2012) Meta-analytic insights into evolutionary ecology: an introduction and synthesis. Evol Ecol 26(5):1085–1099

Nakagawa S, Santos ESA (2012) Methodological issues and advances in biological meta-analysis. Evol Ecol 26(5):1253–1274

Nakagawa S, Schielzeth H (2010) Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. Biol Rev Camb Philos Soc 85(4):935–956. doi:10.1111/j.1469-185X.2010.00141.x

Prokop ZM, Michalczyk L, Drobniak SM, Herdegen M, Radwan J (2012) Metaanalysis suggests choosy females get sexy sons more than "good genes". Evolution 66(9):2665–2673

R Development Core Team (2011) {R}: a language and environment for statistical computing. http://www.R-project.org/

Rafferty NE, Ives AR (2013) Phylogenetic trait-based analyses of ecological networks. Ecology in press

Ross L, Gardner A, Hardy N, West SA (2013a) Ecology, not the genetics of sex determination, determines who helps in eusocial populations. Curr Biol 23(23):2383–2387. doi:10.1016/j.cub.2013.10.013

Ross L, Hardy NB, Okusu A, Normark BB (2013b) Large population size predicts the distribution of asexuality in scale insects. Evolution 67(1):196–206. doi:10.1111/J.1558-5646.2012.01784.X

Rubin DB (1976) Inference and missing data. Biometrika 63(3):581–590

Rubin DB (1987) Multiple imputation for nonresponse in surveys. Wiley, New York, NY

Santos ESA, Nakagawa S (2012) The costs of parental care: a meta-analysis of the trade-off between parental effort and survival in birds. J Evol Biol 25(9):1911–1917. doi:10.1111/J.1420-9101.2012.02569.X

Schielzeth H, Nakagawa S (2013) Nested by design: model fitting and interpretation in a mixed model era. Meth Ecol Evol 4(1):14–24. doi:10.1111/j.2041-210x.2012.00251.x

Sillanpää MJ (2011) On statistical methods for estimating heritability in wild populations. Mol Ecol 20(7):1324–1332. doi:10.1111/j.1365-294X.2011.05021.x

Stone GN, Nee S, Felsenstein J (2011) Controlling for non-independence in comparative analysis of patterns across populations within species. Philos Trans Roy Soc B: Biol Sci 366(1569):1410 –1424. doi:10.1098/rstb.2010.0311

van Buuren S (2012) Flexible imputation of missing data. Chapman and hall/CRC interdisciplinary statistics series. CRC Press, Boca Raton, FL

van de Pol M, Wright J (2009) A simple method for distinguishing within- versus between-subject effects using mixed models. Anim Behav 77(3):753–758. doi:10.1016/j.anbehav.2008.11.006

Waters J, Fraser C, Hewitt G (2013) Founder takes all: density-dependent processes structure biodiversity. Trends Ecol Evol 28(2):78–85. doi:10.1016/j.tree.2012.08.024

Wright S (1934) An analysis of variability in number of digits in an inbred strain of guinea pigs. Genetics 19(6):506–536

Chapter 12
Multimodel-Inference in Comparative Analyses

László Zsolt Garamszegi and Roger Mundry

Abstract Multimodel inference refers to the task of making a generalization from several statistical models that correspond to different biological hypotheses and that vary in the degree of how well they fit the data at hand. Several approaches have been developed for such purpose, and these are widely used, mostly for intraspecific data, i.e., in a non-phylogenetic framework, to draw inference from models that consider different predictor variables in different combinations. Adding the phylogenetic component, in theory, calls for a more extended exploitation of these techniques as several hypotheses about the phylogenetic history of species and about the mode of evolution should also be considered, all of which can be flexibly incorporated and combined with different statistical models. Here, we highlight some biological problems that inherently imply multimodel approaches and show how these problems can be tackled in the phylogenetic generalized least squares (PGLS) modeling framework based on information-theoretic approaches (e.g., by using Akaike's information criterion, AIC) or maximum likelihood. We present a conceptual framework of model selection for phylogenetic comparative analyses, where the goal is to generalize across models that involve different combinations of predictors, phylogenetic hypotheses, parameters describing the mode of evolution, and error structures. Although this overview suggests that a model selection strategy may be useful in several situations, we note that the performance of the approach in the phylogenetic context awaits further evaluation in simulation studies.

L. Z. Garamszegi (✉)
Department of Evolutionary Ecology, Estación Biológica de Doñana-CSIC,
Seville, Spain
e-mail: laszlo.garamszegi@ebd.csic.es

R. Mundry
Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
e-mail: roger_mundry@eva.mpg.de

12.1 Introduction

The world is so complex that researchers are often confronted with the challenge of assessing a large number of biological explanations for a given phenomenon (Chamberlin 1890). Making drawing inference from multiple hypotheses traditionally involves the evaluation of the appropriateness of different statistical models that describe the relationship among the considered variables. This task can be seen as a model selection problem, and there are three general approaches that allow such inference based on statistical analysis. The approach that dominated applied statistics for decades is that of null-hypothesis significance testing (NHST) (Cohen 1994). Applying NHST, one typically states a null-hypothesis of no influence or no difference, which is then rejected or not based on a significance threshold (conventionally, $P = 0.05$ that specifies the probability that one would obtain the observed data given the null hypothesis were true). In this framework, nested multiple models can be examined in a stepwise fashion, in which terms can be eliminated or added based on their significance following a backward or forward process (but see, e.g., Mundry and Nunn 2008; Whittingham et al. 2006; Hegyi and Garamszegi 2011 for problems with stepwise model selection). The second approach is Bayesian inference where one considers a range of 'hypotheses' (e.g., model parameters) and incorporates some prior knowledge about the probability of the particular model parameter values to update one's 'belief' in what are more and less likely model parameters (Congdon 2003; Gamerman and Lopes 2006). Bayesian inference has a long history, but only recent increases in computer power made its application feasible for a wide range of problems (for relevance for comparative studies, see Chaps. 10 and 11). The third, relatively recent approach to statistical inference is based on information theory (IT) (Burnham and Anderson 2002; Johnson and Omland 2004; Stephens et al. 2005). Here, a set of candidate models, which represent different hypotheses, is compared with regard to how well they fit the data. A key component of the IT approach is that the measure of model fit is penalized for model complexity (i.e., the number of estimated parameters), and, as such, IT-based inference aims at identifying models that represent a good compromise between model fit and model complexity. Most frequently, IT-based inference goes beyond simply choosing the best model (out of the set of candidate models) and allows accounting for model selection uncertainty (i.e., the possibility that several models receive similar levels of support from the data).

Although model selection is classically viewed as a solution to the problem caused by the large number of potential combinations of predictors that may affect the response variable, here we propose that the comparative phylogenetic framework involves a range of questions that require multimodel inference and approaches based on IT. In particular, we emphasize that in addition to the variables included in the candidate models, the models can also differ in terms of other parameters that describe the mode of evolution, or account for phylogenetic uncertainty and heterogeneities in sampling effort. In this chapter, we present general strategies for drawing inference from multiple evolutionary models in the framework of

phylogenetic generalized least squares (PGLS). We formulate our suggestions merely on a conceptual basis with the hope that these will stimulate further research that will assess the performance of the methods based on simulations. We envisage that such simulation studies are crucial steps before implementing model selection routines into the practice of phylogenetic modeling. Our discussion is accompanied with an Online Practical Material (hereafter OPM) available at http://www.mpcm-evolution.com, which demonstrates how our methodology can be applied to real data in the *R* statistical environment (R Development Core Team 2013).

12.2   The Fundaments of IT-based Multimodel Inference

Given that a considerable number of primary and secondary resources discuss the details of the IT-based approach (Burnham and Anderson 2002; Claeskens and Hjort 2008; Garamszegi 2011; Konishi and Kitagawa 2008; Massart 2007), we avoid giving an exhausting description here. However, in order to make our subsequent arguments comprehensible for the general readership, we first provide a brief overview on the most important aspects of the approach.

12.2.1   Model Fit

The central idea of an IT-based analysis is to compare the fit of different models in the candidate model set (see below). However, it is trivial that more complex models show better fits (e.g., larger R^2 or smaller deviance). Hence, an IT-based analysis aims at identifying those models (in the set of candidate models; see below) that represent a good compromise between model complexity and model fit, in other words, parsimonious models. Practically, this is achieved by penalizing the fit of the models by their complexity. One way of doing this is to use Akaike's information criterion (AIC), namely

$$\text{AIC} = -2 \ln \mathcal{L}_{(\text{model}|\text{data})} + 2k, \qquad (12.1)$$

where

$\mathcal{L}_{(\text{model}|\text{data})}$ maximum likelihood of the model given the data and the parameter estimates,

k   the number of parameters in the model ($-2 \ln \mathcal{L}_{(\text{model}|\text{data})}$ is known as "deviance").

Two models explaining the data equally well will have the same likelihood, but they might differ in the number of parameters estimated. Then, the model with the smaller number of parameters will reveal the smaller AIC (and the difference in the AIC values of the more complex and the simpler model will be twice the

difference of the numbers of parameters they estimate). Hence, in an IT-based analysis, the model with the smaller AIC is considered to be 'better' because it represents a more parsimonious explanation of the response investigated.[1] Note-worthy, some argue that AIC-based inference can select for overly complex models and suggest alternative information criteria (Link and Barker 2006). Here, we continue focusing on AIC with the notion that the framework can be easily tailored for other metrics.

The core result of an IT-based analysis is a set of AIC values associated with a set of candidate models. However, unlike P values, AIC values do not have an interpretation in themselves but receive meaning only by comparison with AIC values of other models, fitted to the exact same response. The model with the smallest AIC is the 'best' (i.e., best compromise between model parsimony and complexity) in the set of models. However, in contrast to an NHST analysis, it would be misleading to simply select the best model and discard the others. This is because the best model according to AIC (i.e., the one with the smallest AIC) might not be the model that explicitly describes the truth (in fact, it is unlikely to ever be). Such discrepancies can happen for various reasons, including stochasticity in the sampling process (i.e., a sample is used to draw inference about a population), measurement error in the predictors and/or the response, or unknown predictors not being in the model, to mention just a few. An analysis in the framework of a phylogenetic comparative analysis expands this list considerably to include, for instance, imperfect knowledge about the phylogenetic history or the underlying model of evolution (e.g., Brownian motion or Ornstein-Uhlenbeck). An IT-based analysis allows dealing elegantly with such model selection uncertainty by explicitly taking it into account (see below).

12.2.2 Candidate Model Set

A key component in an IT-based analysis is the candidate set of models to be investigated, which classically includes models with different combinations of predictors. The validity of the analysis is conditional on this set, and if the candidate model set is not a reasonable one, the results will be deceiving (Burnham and Anderson 2002; Burnham et al. 2011). Hence, the development of the candidate set needs much care and is a crucial and potentially challenging step of an IT-based analysis. First of all, different models might represent different research hypotheses. For instance, one might hypothesize that brain size might have co-evolved with social complexity (e.g., group size), ecological complexity (e.g., seasonality in food availability), or both. However, in biology, it is frequently not easy to come up with such a clearly defined set of potentially competing models,

---

[1] When drawing inference in an IT framework, it is essential to not mix it up with the NHST framework. Most crucially, it does not make sense to select the best model based on AIC and then test its significance or the significance of the predictors it includes.

and hence one frequently sees candidate model sets that encompass all possible models that can be built out of a set of predictors. Furthermore, in the context of phylogenetic comparative analysis, different models in the candidate set might represent different evolutionary models (e.g., Brownian motion or an Ornstein-Uhlenbeck process) or different phylogenies. It is important to emphasize that in a phylogenetic comparative analysis both these aspects (and also other ones) can be reflected in a single candidate set of models; that is, the candidate set might comprise models that represent combinations of hypotheses about the coevolution of traits, the model of evolution, and the phylogenetic history.

12.2.3 Accounting for Model Uncertainty

There are several ways of dealing with model selection uncertainty (i.e., with the fact that not only one model is unanimously selected as best). One way is to consider Akaike weights. Akaike weights are calculated for each model in the set and can be thought of as the probability of the actual model to be the best in the set of models (although there are warnings against such interpretations, e.g., see Bolker 2007). From Akaike weights, one can also derive the evidence ratio of two models, which is the quotient of their Akaike weights and tells how much more likely is one of the two models (i.e., the one in the numerator of the evidence ratio) to be the best model. Akaike weights can also be used to infer about the importance of individual predictors by summing Akaike weights for all models that contain a given predictor. The summed Akaike weight for a given predictor then can be considered analogous to the probability of it being in the best model of the set (see also Burnham et al. 2011; Symonds and Moussalli 2011).

12.2.3.1 Model Averaging

One can also use Akaike weights for model averaging of the estimated coefficients associated with the different predictors. Here, the estimated coefficients (e.g., regression slopes) are averaged across all models (or across a confidence set of best models[2]) weighted by the Akaike weights of the corresponding models (see also Burnham et al. 2011; Symonds and Moussalli 2011). Hence, an estimate of a coefficient from a model having a large Akaike weight contributes more to the

---

[2] Another way of dealing with model selection uncertainty is to consider the best model confidence set, which contains the models that can be considered as best with some certainty. Different criteria do exist to identify the best model confidence set among which the most popular are to include those models that differ in AIC from the best model by at most some threshold (e.g., 2 or 10) or, alternatively, to include those models for which their summed cumulative Akaike weights (from largest to smallest) just exceed 0.95. In this chapter, we do not consider such subjective thresholds further, and throughout the remaining discussion we refer to model averaging in a sense that it is made across the full model set.

averaged value of the coefficient. When using model averaging of the estimated coefficients, there are two ways of treating models in which a given predictor is not present: one is to simply ignore them (the 'natural' method), and one is to set the estimated coefficient to zero for models in which the given predictor is not included (the 'zero' method; Burnham and Anderson 2002; Nakagawa and Hauber 2011). Using the latter penalizes the estimated coefficient when it is mainly included in models with low Akaike weights, and to us, this seems to be the better method.

12.3 Model Selection Problems in Phylogenetic Comparative Analyses

There can be several biological questions involving phylogenies, which necessitate inference from more than one model that are equally plausible hypothetically. Most readers might have encountered such a challenge when judging the importance of different combinations of predictor variables. However, in addition to parameters that estimate the effects of different predictors, in a phylogenetic model, there are several other parameters that deal with the role of phylogenetic history or with another error term (e.g., within-species variance). The statistical modeling of these additional parameters often requires multiple models that differently combine them, even at the same set of predictors. Below, we demonstrate that in most of these situations the observer is left with the classical problem of model selection, when s/he needs to draw inferences from a pool of models based on their fit to the data. Accordingly, the same general framework can be applied: Competing biological questions are first translated into statistical models, and then, multimodel inference is used for generalization.

12.3.1 Selecting Among Evolutionary Models with Different Combinations of Predictors

The classical problem of finding the most plausible combination of predictors to explain interspecific variation in the response variable while accounting for the phylogeny of species is well exemplified in the comparative literature. Starting from a pioneering study by Legendre et al. (1994), a good number of studies exist that evaluate multiple competing models to assess their relative explanatory value and to draw inferences about the effects of particular predictors. Below, as an appetizer, we provide summaries of two of these studies to demonstrate the diversity of questions that can be addressed by using the model selection framework. In the OPM, we give the *R* code that can be easily tailored to any biological problem requiring an AIC-based information-theoretic approach.

Terribile et al. (2009) investigated the role of four environmental hypotheses mediating interspecific variation in body size in two snake clades. These hypotheses

emphasized the role of heat balance as given by the surface area-to-volume ratio, which in ectothermic vertebrates may influence heat conservation (e.g., small-bodied animals may benefit from rapid heating in cooler climates), habitat availability (habitat zonation across mountains limits habitat areas that ultimately select for smaller species), primary productivity (low food availability can reduce growth rate and delay sexual maturity, which would in turn result in small-bodied species in areas with low productivity), and seasonality (large-bodied species may be more efficient in adapting to seasonally fluctuating resources that often include periods of starvation). To test among these hypotheses, the authors estimated the extent to which the patterns of body size are driven by current environmental conditions as reflected by mean annual temperature, annual precipitation, primary productivity, and range in elevation. They challenged a large number of models with data and chose the best model that offered the highest fit relative to model complexity to draw inference about the relative importance of different hypotheses. This best model included all main predictors, but the amount of variation explained differed between Viperidae and Elapidae, the two snake clades investigated. Moreover, the relative importance for each predictor also varied, as indicated by the summed Akaike weights. Consequently, none of the proposed hypotheses was overwhelmingly supported or could be rejected, and the mechanisms constraining body size in snakes can even vary from one taxonomic group to another.

A recent phylogenetic comparative analysis of mammals focused on the determinants of dispersal distance, a variable of major importance for many ecological and evolutionary processes (Whitmee and Orme 2013). Dispersal distance can be hypothesized as a trait being influenced by several constraints arising from life history, a situation that necessitates multipredictor approaches. For example, larger body size can allow longer dispersal distances because locomotion is energetically less demanding for larger-bodied animals. Second, home range size may be important, as dispersing individuals of species using larger home ranges may need to move longer distances to find empty territories. Furthermore, trophic level, reflecting the distribution of resources, may mediate dispersal distance with carnivores requiring more dispersed resources than herbivores or omnivores. Intraspecific competition may also affect dispersal: species maintaining higher local densities may also show higher frequencies of distantly dispersing individuals which thereby encounter less competition. Finally, investment in parental activities can be predicted to negatively influence dispersal, as species that wean late and mature slowly will create less competitive conditions for their offspring than species with fast reproduction. To simultaneously evaluate the plausibility of these predictors, Whitmee and Orme (2013) applied a model selection strategy based on the evaluation of a large number of models composed of the different combinations (including their quadratic terms) of the considered predictors. Even the best-supported multipredictor models had low Akaike weights, indicating no overwhelming support for any particular model. Therefore, they applied model averaging to determine the explanatory role of particular variables, which indicated that home range size, geographic range size, and body mass are the most important terms across models.

12.3.2 Dealing with Phylogenetic Uncertainty: Inference Across Models Considering Different Phylogenetic Hypotheses

While phylogenetic comparative studies necessarily require a phylogenetic tree, the true phylogeny is never known and must be estimated from morphological or, more recently, from genetic data; thus, phylogenies always contain some uncertainty (see detailed discussion in Chap. 2). In several cases, more than one phylogenetic hypothesis (i.e., tree) can be envisaged for a given set of species, and it might be desirable to test whether the results found for a given phylogenetic tree are also apparent for other, similarly likely trees.

With GenBank data and nucleotide sequences for phylogenetic inference, the above problem is not restricted anymore to the comparison of a handful of alternative trees corresponding to different markers. Nonetheless, the reconstruction of phylogenies from the same molecular data still raises uncertainty issues at several levels. Different substitution models and multiple mechanisms can be considered for sequence evolution, each leading to different sets of phylogenies that can be considered (note that this is also a model selection problem). Moreover, even the same substitution model can lead to various phylogenetic hypotheses with similar likelihoods. As a result, in the recent day's routine, several hundreds or even thousands of phylogenetic trees are often available for the same list of species used in a comparative study. The most common way to deal with such a large sample of trees is the use of a single, consensus tree in the phylogenetic analysis. However, although this approximation is convenient from a practical perspective, using an 'average' tree does not capture the essence of uncertainty, which lies in the variation across the trees. The whole sample of similarly likely trees defines a confidence range around the phylogenetic hypothesis (de Villemereuil et al. 2012; Pagel et al. 2004).

For the appropriate treatment of phylogenetic uncertainty, one needs to incorporate an error component that is embedded in the pool of trees that can be envisaged for the species at hand. Martins and Hansen (1997) proposed that most questions in relation to the evolution of phenotypic traits across species can be translated into the same general linear model:

$$\mathbf{y} = \beta \mathbf{X} + \varepsilon, \tag{12.2}$$

where
- **y**   is a vector of characters or functions of character states for extant or ancestral taxa,
- **X**   is a matrix of states of other characters, environmental variables, phylogenetic distances, or a combination of these,
- **β**   is a vector of regression slopes,
- **ε**   is a vector of error terms with an assumed structure.

ε is composed of at least three types of errors that can be assembled in a complex way: ε_S, the error due to common ancestry; ε_M, the error due to within-species variance or measurement error; and ε_P, the error due phylogenetic uncertainty. The regression technique based on PGLS when combined with maximum likelihood (ML) model fitting offers a flexible way to handle and combine the errors ε_S and ε_M (for example, they can be treated additively if they are independent, see Chaps. 5 and 7). However, simultaneously handling the third error, the one that is caused by phylogenetic uncertainty, ε_P, is more challenging, because it is not an independent and additive term (Martins 1996). Approaches based on Bayesian sampling that are discussed in Chap. 10 offer a potential solution. They allow the use of a large number of similarly likely phylogenetic trees by effectively weighting parameter estimates across their probability distribution and can also incorporate errors due to within-species variance (de Villemereuil et al. 2012). However, widely available Bayesian methods can be sensitive to prior settings and are not yet implemented in the commonly used statistical packages.

We propose a simpler solution and suggest that when combined with multi-model inference, approaches based on PGLS can be used to deal with uncertainties in the phylogenetic hypothesis. The underlying philosophy of this approach is that when a list of trees is available, each of them can be used to fit the same model describing the relationship between traits using ML. Subsequently, parameter estimates (e.g., intercepts and slopes) can be obtained from the resulting models, which can then be averaged with a weight that is proportional to the relative fit of the corresponding model to the data. The output will not only provide a single average effect (as is the case when using a single model fitted to a consensus tree) but will also include a confidence or error range as obtained from the variance of model parameters across models associated with different trees. This interval can be interpreted as a consequence of the uncertainty in the phylogenetic hypothesis, that is, the mean estimate (model-averaged slope, or the slope that is based on the consensus tree) with the associated uncertainty component (variance among particular slopes) will form the results together. The logic of analyzing the inter-specific data on each possible phylogeny to obtain a sample of estimates and then to calculate summary statistics from this distribution was already proposed by Martins (1996). Our favored method differs with regard to that it applies a model-averaging technique to derive the mean and confidence interval from the frequency distribution of parameters. This can be important, because if the pool of the trees across which the models are fitted reflects the likelihood of particular trees explaining the evolution of taxa, the resulting model-averaged parameter estimates will also reflect this variation.

Although apparently different trees are used in each model, drawing inference across them does not violate the fundaments of information theory that assumes that each model is fitted to the very same data. Different trees can be regarded as different hypotheses that arise from identical nucleotide sequence information. They are actually just different statistical translations of the same biological information and act like scaling parameters on the tree. The approach may be particularly useful when a large number of alternative trees are at hand (e.g., in the

form of a Bayesian sample originating from the same sequence data). When only a handful of phylogenies is available (e.g., from other published papers), model-averaged means and variances can also be calculated, but conclusions would be conditional on the phylogenies considered (i.e., some alternative phylogenies may have not been evaluated). Furthermore, fitting models to trees that correspond to different marker genes calls for philosophical issues about the underlying assumption concerning the use of the same data.

In Fig. 12.1, we illustrate how our proposed model averaging works in practice (the underlying computer codes are available in the OPM). In this example, we tested for the evolutionary relationship between brain size and body size in primates by using PGLS regression methods with ML estimation of parameters. We considered a sample of reasonable phylogenetic hypotheses in the form of 1,000 trees as obtained from the *10KTrees Project* (Arnold et al. 2010). When using the consensus tree from this tree sample, we can estimate that the phylogenetically corrected allometric slope is 0.287 (*SE* = 0.039, solid line in Fig. 12.1). However, using different trees from *10KTrees* pool in the model provides slightly different results for the phylogenetic relationship between traits, as the obtained slopes vary (gray lines in the left panel of Fig. 12.1). The model-averaged regression slope yields 0.292 (model-averaged *SE* = 0.041, dashed line in the left panel of Fig. 12.1). This mean estimate is quite close to what one can obtain based on the consensus tree, but the variation between the particular slopes corresponding to different trees in the sample delineates some uncertainty around the averaged allometric coefficient. Few models in the ML sample provide extreme estimates (note that, model fitting with one particular tree even results in a negative slope, left panel of Fig. 12.1). However, these models were characterized by a very poor model fit; thus, their potential influence is downweighted in the model-averaged mean estimate.

The benefit of using the AIC-based method to account for phylogenetic uncertainty over Bayesian approaches is that the former does not require prior information on model parameters that would affect the posterior distribution of parameters, an issue that is often challenging in the Bayesian framework (Congdon 2006) and that is also demonstrated in Fig. 12.1. In the right panel, we applied Markov chain Monte Carlo (MCMC) procedure to estimate the posterior distribution of parameter values from the same PGLS equation by using (Pagel et al. 2004; Pagel and Meade 2006) *BayesTraits* with the same interspecific data and pool of trees (see also Chap. 10). Supposing that we have no information to make an expectation about the range where parameter estimates should fall, we are constrained to use flat and uniform prior distributions (e.g., spanning from −100 to 100).[3] When we used MCMC to

---

[3] It may not be necessarily applied to the current biological example, because allometric regressions are intensively studied (e.g., Bennett and Harvey 1985; Hutcheon et al. 2002; Iwaniuk et al. 2004; Garamszegi et al. 2002). Therefore, results from a large number of studies on other vertebrate taxa may be used to define a narrower and more informative prior. However, in this example simulated on the general situation when no preceding information on the expected relationship is available. Note that technically *BayesTraits* only allows uniform priors for continuous data.
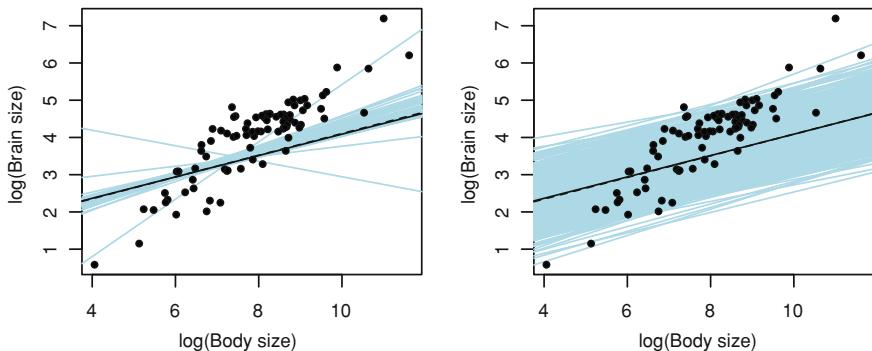
Fig. 12.1 Estimated regression lines for the correlated evolution of two traits (body size and brain size in primates) when different hypotheses for the phylogenetic relationships of species are considered and when ML (*left panel*) or MCMC (*right panel*) estimation methods are used in the AIC-based or Bayesian framework, respectively. *Gray lines* show the regression slopes that can be obtained for alternative phylogenetic trees (*left panel* 1,000 ML models fitted to different trees, *right panel* 1,000 models that the MCMC visited in the Bayesian framework). The alternative trees originate from a sample of 1,000 similarly likely trees that can be proposed for the same nucleotide sequence data (Arnold et al. 2010). The *dashed bold line* represents the slope estimate that can be derived by model averaging over the particular ML estimates (*left panel*) or by taking the mean of the posterior distribution from the MCMC sample of 1,000 models (*right panel*). Both methods provide a mean estimate over the entire pool of trees by incorporating the uncertainty in the underlying phylogenetic hypothesis. The *solid bold line* shows the regression line that can be fitted when the single consensus tree is used. The model-averaged slope, the mean of the posterior distribution, and the one that corresponds to the consensus highly overlap in this example (which may not necessarily be the case). However, the precision by which the mean can be estimated is different between ML and MCMC approaches, as the latter introduces a larger variance in the slopes in the posterior sample

sample from a large number of models with different parameters and trees and took 1,000 estimates from the posterior distribution of slopes, we detected that the estimate is accompanied by a considerable uncertainty (Fig. 12.1, right panel). For comparison, the 95 % confidence interval of the allometric coefficients obtained from the ML sample is 0.278–0.312, while it is 0.211–0.373 for the MCMC sample (i.e. the confidence interval obtained from the Bayesian framework is almost five times wider than that from the AIC-based inference). Consequently, the Bayesian approach introduces an unnecessary uncertainty due to the dominance of the prior distribution on the posterior distribution.

Another benefit of using ML model fitting over a range of phylogenetic hypotheses in conjunction with model averaging is that by doing so we can exploit the flexibility of the PGLS framework. For example, as we discussed above, one can evaluate different sets of predictor variables when defining models, or as we explain below, one can also take into account additional error structures (e.g., due to within-species variation) or different models of trait evolution (e.g., Brownian motion or an Ornstein-Uhlenbeck process). These different scenarios can be simultaneously considered during model definition, but can also be combined with

alternative phylogenetic trees (some examples are given in the OPM). This will result in a large number of candidate models representing different evolutionary hypotheses, over which model averaging may offer interpretable inference.

Box 12.1 A simulation strategy for testing the performance of multimodel inference

The behaviour of the AIC-based framework to account for phylogenetic uncertainty requires simulation studies that consist of the following steps. First, one needs to simulate a tree for a considered number of species and under some scenario for the underlying model (e.g., time-dependent birth–death model or just a random tree). The next step is then to simulate species-specific trait data along the branches of the generated phylogeny. To obtain simulated tip values, we also need to consider a model to describe the evolutionary mechanism in effect (e.g., Brownian motion or an OU process). We might also consider other constraints for trait evolution, for example, by defining a correlation structure (a zero or a nonzero covariance) for two coevolving traits. These parameters will serve as generating values, and the underlying tree and the considered covariance structure will reflect the truth that we want to recover in the simulation. If the interest is to examine the performance of the model-averaging strategy to account for phylogenetic uncertainty, we need to generate a sample of trees that integrates a given amount of variance (e.g., both the topology and branch lengths are allowed to vary to some pre-defined degree). For each simulation, we can then fit a model estimating the association between the two traits by controlling for phylogenetic effects. The phylogeny used in this model to define the expected variance–covariance structure on the one hand can be the consensus tree calculated for the whole sample of trees. On the other hand, we can also fit the model to each tree in the sample and then do a model averaging to obtain an overall estimate for the parameter of interest (e.g., slope or correlation as calculated from the model). By simulating new trait data (and optionally new pools of trees), we can repeat the whole process a large number of times (i.e., 1,000 or 10,000 times). At each iteration, we will, hence, obtain estimates (either over the consensus tree or over the entire sample of trees through model averaging) for the parameter of interest. Finally, we can compare the distribution of these parameters over simulations with the generating parameter state. The difference between the mean of the distribution and the generating value will inform about bias of the approach, while the width of the distribution informs about precision (the uncertainty in parameter estimation).

As an important cautionary note, we emphasize that the performance of the AIC-based method based on model averaging still requires further assessment with both simulated and empirical data. In Box 12.1, we describe the philosophy of an

appropriate simulation study that can efficiently test the performance of averaging parameters over a large number of models corresponding to different hypotheses about phylogenies or other evolutionary patterns.

12.3.3 Variation Within Species

One of the advantages of the PGLS approach is that it allows accounting for within-species variation, which broadly includes true individual-to-individual or population-to-population variation, and also other sources of variation in the estimates of taxon trait values such as measurement error (see Ives et al. 2007; Hansen and Bartoszek 2012; and Chap. 7). Given that these different sources of error can be translated into different models, selecting among these may also be performed by model selection. Does a model that considers within-species variation perform better than a model that neglects such variation? Such simple questions can be developed further as by applying the general Eq. 12.2, in which different error structures (e.g., phylogenetic errors and measurement errors, or measurement error on one trait may correlate with measurement error on another trait) can be combined in different ways.

For example, when considering intraspecific variation in an interspecific context, we can evaluate at least four models and compare them based on their relative fit (here, we are only focusing on the main logic; for details on how to take into account intraspecific variation, see Chap. 7). First, as a null model, we can fit a model that is defined as an ordinary least squares regression (i.e., with a covariance matrix for the residuals based on a star phylogeny and measurement errors being zero). Then, we can investigate a model that does account for phylogeny but not for the uncertainty in the species-specific trait values (conditioned on the true phylogeny, while measurement errors are assumed to be equal to zero), and also a model that considers measurement error but ignores the phylogenetic structure (unequal and nonzero values along the diagonal of the measurement error matrix, and a phylogenetic covariance matrix representing a star phylogeny). Finally, we wish to test a model that includes both error structures (the joint variance–covariance matrix reflecting the phylogeny and the known measurement errors). To obtain parameter estimates and to make appropriate evolutionary conclusions, the observer can rely on the model that offers the best fit to the data as indicated by the corresponding AIC (but only if one model is unanimously supported over the others). Such a simple model selection strategy can be followed in the OPM of Chap. 7. Note that for the appropriate calculation of AIC according to Eq. 12.1 (and thus for the meaningful comparison of models), it is required that the number of estimated parameters is determined, which may be difficult when parameters in both the mean and variance components are estimated. This problem can be avoided by a smart definition of models (e.g., by defining analog models that estimate the same number of parameters even if these are known to be zero). In any case, the approach requires further validation by simulation studies.

Methods that account for within-species variation can also deal with a situation, in which different sample sizes (n) are available for different species, implying that data quality might be heterogeneous (i.e., larger errors in taxa with lower sampling effort; see Chap. 7 for more details). For example, if within-species variances or standard errors are unknown, one can fit a measurement error model by using $1/n$ as an approximation of within-species variance.

Another way to incorporate heterogeneous sampling effort across species into the comparative analyses is to apply statistical weights in the model. A particular issue arising in this case is that weighting can result in a large number of models (with potentially different results). For example, by using the number of individuals sampled per taxon as statistical weights in the analysis, we enforce weights differing a lot between species that are already sampled with sufficient intensity (e.g., the underlying sample size is 20 at least) but still differ in the background research effort (e.g., 100 individuals are available for one species, while 1,000 for another). However, if we log- or square-root-transform within-species sample sizes and use these as weights, more emphasis will be given on differences between lightly sampled species than on differences between heavily sampled species. Continuing this logic, and applying the appropriate transformation, we can create a full gradient that scales differences in within-species sample sizes along a continuum spanning from no differences to large differences between species with different within-species sample sizes.

For illustrative purposes (Fig. 12.2, left panel), we have created such a gradient of statistical weights by the combination of the original species-specific sample sizes (n) and an emphasis parameter (the 'weight of weights') that we will label ω; ω is simply an elevation factor that ranges from 1 to 1/100 and defines the exponent of n. If ω is 1, the original sample sizes are used as weights in the analysis. If ω is $1/2 = 0.5$, the square-root-transformed values serve as weights, and differences between small sample sizes become more emphasized than differences between species with larger sample sizes. $\omega = 1/\infty = 0$ represents the scenario in which all species are considered with equal weight ($n^0 = 1$), so the model actually represents a model that does not take into account heterogeneity in sampling effort. Other transformations on sample sizes based on different scaling factors that create a gradient can also be envisaged.

Using the parameter ω, we provide an example for the study of brain size evolution based on the allometric relationship with body size (Fig. 12.2 right panel, the associated R code is provided in the OPM). We have created a set of phylogenetic models that also included statistical weights in the form of the ω exponent of within-species sample sizes. The scaling factor ω varied from 0 to 1. We challenged these models with exactly the same data using ML; thus, model fit statistics (e.g., AIC) are comparable. We found that when accounting for phylogenetic relationships, the $\omega = 0$ scenario provides by far the best fit, implying that weighting species based on sample size is not important. This finding is not surprising, given that both traits, brain size and body size, have very high repeatability ($R > 0.8$). Thus, relatively few individuals provide reliable information on the species-specific trait values. Giving different weights to different
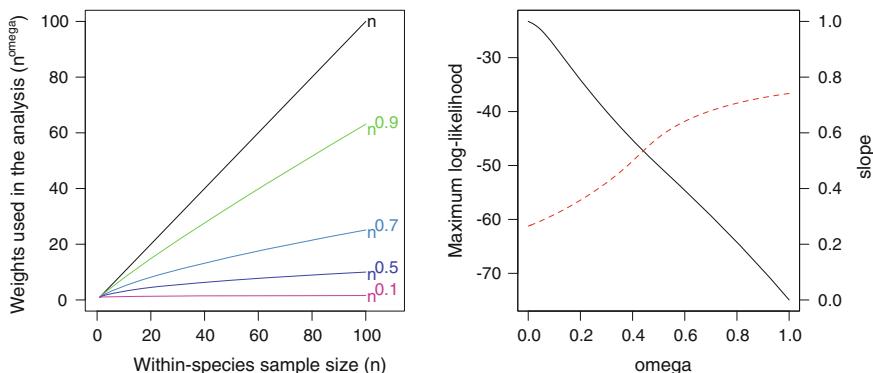
Fig. 12.2 The effect of using different transformations of the number of individuals as statistical weights. The *left panel* shows how differences between species are scaled when the underlying within-species sample sizes are transformed by exponentiating them with the exponent ω. ω varied between 1 (untransformed sample sizes maximally emphasizing differences in data quality between species) and 0 (all species have the same weight; thus, data quality is considered to be homogeneous). The *right panel* shows the maximized log-likelihood (*black solid line*) and the estimated slope parameters of models (*red dashed line*) for the brain size/body size evolution that implement weights that are differently scaled by ω

species based on the underlying sample sizes would actually be misleading; the use of different ω values leads to qualitatively different parameter estimates for the slope of interest (Fig. 12.2, right panel). This indicates that the results and conclusions are highly sensitive to how differences in sampling effort are treated in the analysis. Note that the above exercise only makes sense if (1) there is a considerable variation in within-species sample size and (2) if there is no phylogenetic signal in sample sizes. These assumptions require some diagnostics prior to the core phylogenetic analysis (see an example in Garamszegi and Møller 2012).

Garamszegi and Møller (2007) relied on a similar approach in a study of the ecological determinants of the prevalence of the low pathogenic subtypes of avian influenza in a phylogenetic comparative context. It was evident that there was a vast variance in sampling effort across species, as within-species sample size varied between 107 and 15,657. Therefore, when assessing the importance of the considered predictors, it seemed unavoidable to simultaneously account for common ancestry and heterogeneity in data quality. The application of the strategy of scaling the weight factor yielded that, contrary to the above example, the highest ML was achieved by a certain combination of the weight and phylogenetic scaling parameters. That finding was probably driven by the relatively modest repeatability of the focal trait (prevalence of avian influenza), suggesting that, due to different sample sizes, data quality truly differed among species.

We advocate that the importance of a correction for sample size differences between species is an empirical issue that can vary from data to data, which could (and should) be evaluated. We provided a strategy by which the optimal scaling of weight factors can be determined. In these examples, an unambiguous support could

be obtained for a single parameter combination. However, we can imagine situations, in which more than one model offers relatively good fit to the data, in which case inference would be better made based on model averaging (corresponding codes are given in the OPM) instead of focusing on a single parameter combination. Furthermore, the evaluation of the sample size scaling factor (as well as the assessment of within-species variance) can be combined with the evaluation of alternative phylogenetic hypotheses, as the IT-based framework offers a potential for the exploration of a multidimensional parameter space. Accordingly, each scaling factor can be incorporated into various models considering different phylogenies (or each phylogenetic tree can be evaluated along a range of scaling factors), and the model selection or model-averaging routines may be used for drawing inference from the resulting large number of models. Again, the performance of these methods necessitates further investigations by simulation approaches.

12.3.4 Dealing with Models of Evolution

12.3.4.1 Comparison of Models for Different Evolutionary Processes

Several phylogenetic comparative methods (e.g., phylogenetic autocorrelation, independent contrasts, and PGLS) assume that the model of trait evolution can be described by a Brownian motion (BM) random-walk process. However, this assumption might be violated in certain cases, and other models might need to be considered. For example, a model based on the Ornstein-Uhlenbeck (OU) process is another choice that takes into account stabilizing selection toward a single or multiple adaptive optima (Butler and King 2004; Hansen 1997; see also discussion in Chaps. 14 and 15). Other model variants of the BM or OU models, such as the model for accelerating/decelerating evolution (AC/DC, Blomberg et al. 2003) or the model for a single stationary peak (SSP, Harmon et al. 2010), can also be envisaged.

  Given that we usually do not have prior information about the 'true' model of evolution, alternative hypotheses about how traits evolved could be considered in statistical modeling. If the considered evolutionary models are mathematically tractable (there are cases when they are not! see Kutsukake and Innan 2013), they can be translated into statistical models suitable for a model selection framework. Accordingly, each model can be fitted to the data, and once finding the one that offers the highest explanatory power, it can be used for making evolutionary inferences. This does not only control for phylogenetic relationships, but knowing which is the most likely evolutionary model can give insight about the strength, direction, and history of evolution acting on different taxa. Importantly, when using a model selection strategy in this context, the observer aims at identifying the single best model that accounts for the mode of evolution; thus, model averaging may not make sense. Therefore, for making robust conclusions, we need to obtain results in which models are well separated based on their AIC in a way that

one model reveals overwhelming support as compared to the others. Alternatively, one could use model averaging to estimate regression parameters (and also to estimate the parameters of the evolutionary model if parameters of different models are analogous), thus accounting for the uncertainty in the assessment of the underlying evolutionary process.

To demonstrate the use of model selection to choose among different evolutionary models, we provide an example from Butler and King (2004), but other illustrative analyses are also available in the literature (Collar et al. 2009, 2011; Harmon et al. 2010; Hunt 2006; Lajeunesse 2009; Scales et al. 2009). Butler and King (2004) re-examined character displacement in *Anolis* lizards on the Lesser Antilles, where lizards live either in sympatry or in allopatry. Where two species coexist, these differ substantially in size, while on islands that are inhabited by only one species, lizards are of intermediate size. Therefore, one can hypothesize that body size differences on sympatric islands result from character displacement (i.e., when two intermediate-sized species came into contact with one another when colonizing an island, they subsequently diverged into a different direction). This hypothesis can be evaluated using alternative models of body size evolution that differ in the degree of how they incorporate processes due to directional selection and character displacement. The authors, therefore, evaluated five different models: (1) BM; (2) an OU process with a single optimum; (3) an OU process with three optima corresponding to large, intermediate, and small body size; (4) another OU model that includes an additional parameter to the three-optima model to deal with the adaptive regimes occurring on the internal branches as an estimable ancestral state; and (5) a model implementing a linear parsimony reconstruction of the colonization events (arrival history of species on the islands). Only the last model assumes character displacement. These models were compared by different methods including AIC, a Bayesian (Schwarz's) information criterion (SIC), and likelihood ratio tests that unanimously revealed that the best-fitting model was the OU model with the reconstructed colonization events (Fig. 12.3). Altogether, the results support the hypothesis that character displacement had an effect on the evolution of body size in *Anolis* lizards that colonized the Lesser Antilles.

12.3.4.2 Parameterization of Models

Another way to cope with the mode of evolution and to improve the fit of any model can be achieved by the appropriate setting of parameters that describe the fine details of the evolutionary process. For example, BM models can be adjusted using the parameters κ, δ, or λ that apply different branch-length transformations on the phylogeny (e.g., κ stretches or compresses phylogenetic branch lengths and thus can be used to model trait evolution along a gradient from punctuational to gradual evolution, while δ scales overall path lengths in the phylogeny and thus can be used to characterize the tempo of evolution) or that assess the contribution of the phylogeny (λ weakens or strengthens the phylogenetic signal in the data)

	BM	**OU(1)**	**OU(3)**	**OU(4)**	**OU(LP)**
$-2\ln\mathcal{L}$	−34.66	−34.66	−40.21	−47.22	−49.69
AIC	−30.66	−26.66	−28.21	−33.22	−37.69
LR		0	5.55	12.56	15.03
P		1	0.24	0.028	0.0046
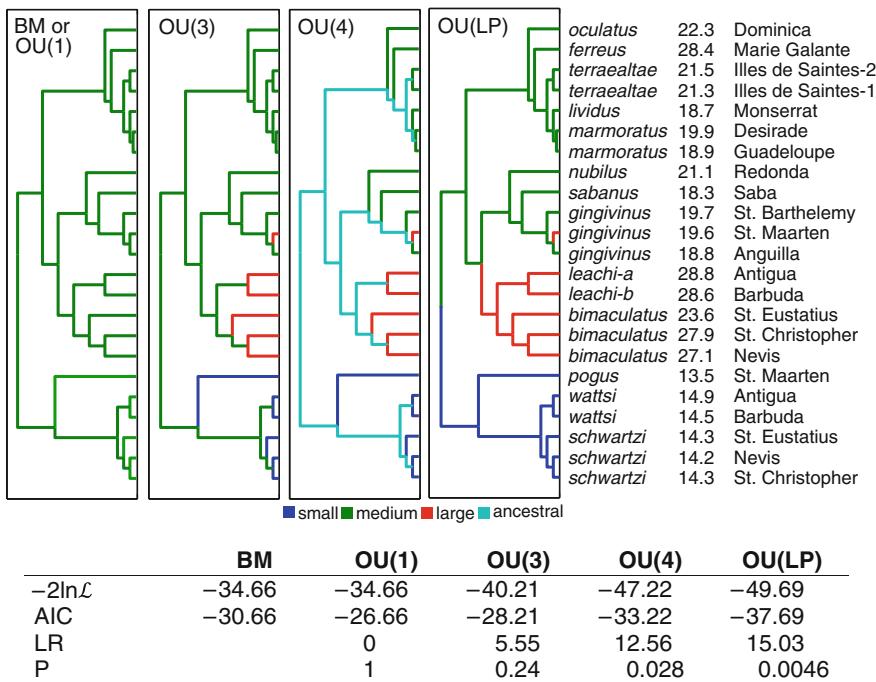
Fig. 12.3 Graphical representation of five evolutionary models considered for the evolution of body size in Anolis lizards inhabiting the islands of Lesser Antilles. *BM* Brownian model; *UO(1)* Ornstein-Uhlenbeck process with a single optimum; *OU(3)*, *OU(4)* Ornstein-Uhlenbeck process with three or four optima, respectively i.g., Ornstein-Uhlenbeck process with four optima, one of which is an ancestral state; *OU(LP)* Ornstein-Uhlenbeck process with implementing a linear parsimony reconstruction of the colonization events, which thus considers character displacement. The table shows the model fit statistics of different models.: deviance, Akaike's information criterion (see Eq. 12.1), and likelihood ratio test comparing the given model with the BM model (LR and the associated *P* values). Modified from Butler and King (2004) with the permission of the authors and University of Chicago Press

(Pagel 1999). Furthermore, the importance of the rates of evolutionary change in character states can also be assessed via estimation of the corresponding parameter (Collar et al. 2009; O'Meara et al. 2006; Thomas et al. 2006). Finally, UO models also operate with particular parameters (such as α for the strength of selection and θ for the optimum) that can take different values (Butler and King 2004; Hansen 1997, see also discussion in Chaps. 14 and 15).

The parameterization of models is a task that requires the investigator to choose among alternative models with different parameter settings, which is typically a model selection problem. This task is usually addressed with likelihood ratio tests, in which a null model (e.g., with a parameter set to be zero) is contrasted with an

alternative model (e.g., with a parameter set to a nonzero value). If the test turns out significant, the alternative model is accepted and used for further analyses (e.g., tests for correlations between traits) and for making evolutionary implications. Another strategy is to evaluate the ML surface of the parameter space and then set the parameter to the value where it reveals the maximum likelihood (i.e., the strategy that most PGLS methods apply). Furthermore, AIC-based information-theoretic approaches can be used to obtain the parameter combinations that offer the best fit to the data.[4]

However, such a best model approach is not always straightforward. Parameter states can span a continuous scale, and it is possible that a broad range of parameter values are similarly likely. For example, the optimal phylogenetic scaling parameter λ is usually estimated using maximum likelihood. This estimation might be robust if the peak of the likelihood surface is well defined (i.e., few parameter states in a narrow range have a very high likelihood, while the remaining spectrum falls into a small likelihood region, Fig. 12.4, upper panels). Our experience, however, is that the likelihood surfaces are rather flat and vary considerably if single species are added or removed from the analysis (especially at modest interspecific sample sizes, Freckleton et al. 2002). This means that a broad range of parameter values describe the data similarly well (Fig. 12.4, lower panels), thus arbitrarily choosing a single parameter value on a flattish surface for further analysis may be deceiving.

We suggest that such uncertainty in parameter estimation can easily be incorporated using model averaging. Applying the philosophy that we followed for dealing with multiple trees or scenarios for the correction for heterogeneous data quality, we can also estimate the parameters of interest (e.g., ancestral state, slope, or correlation between two traits) at a wide range of the settings of the evolutionary parameters. Given that IT-based approaches typically compare sets of discrete models, we need to create a large number of categories for the continuous parameter (e.g., by defining a finite number, such as 100 or 1,000, bins for λ in increasing order between the interval of 0–1) that can be used to condition different models. Then, inference across this large number of models based on their relative fit to the data can be made, and given that intermediate states between the large number of categories are meaningful, interpretations can be extended to a continuous scale. Therefore, evolutionary conclusions can be formulated based on the parameter estimates that are averaged across models receiving different levels of support instead of obtaining them from a single model. In theory, λ can be model averaged as well, but when the maximum likelihood surface is flat (meaning that many models with different λs will have similar AIC), deriving a single mean estimate may be misleading. In such a situation, only estimates together with their model-averaged standard errors (or confidence intervals) make sense.

In the OPM, we show for λ how this model averaging works in practice. We also provide examples for the case when the exercise for model parameters is

---

[4] As long as the number of parameters is equal, AIC and ML reveal the same.
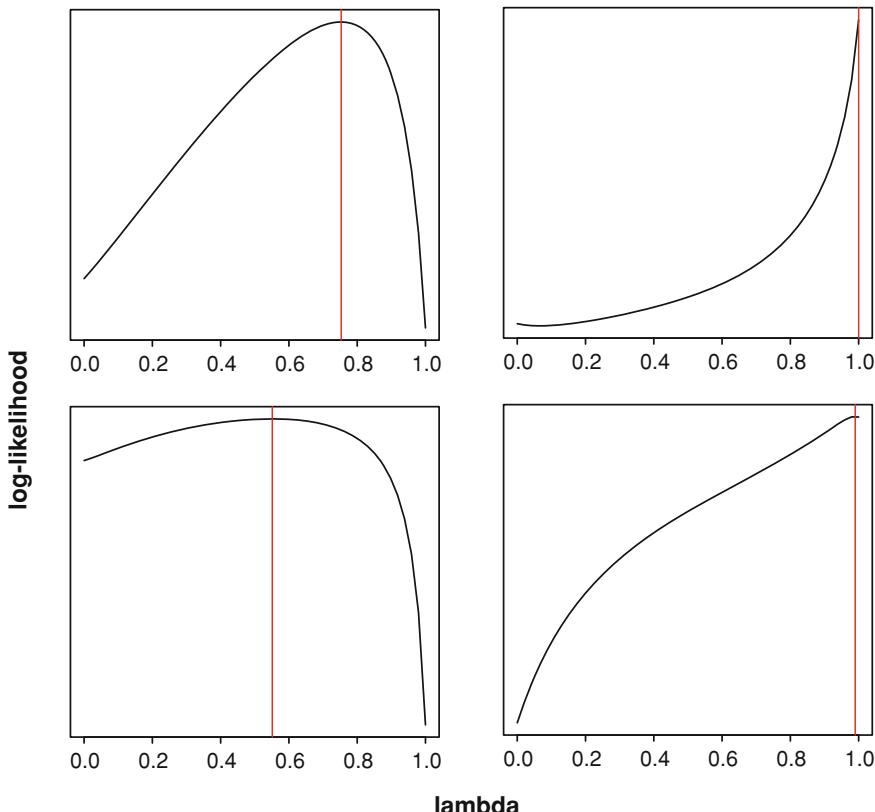
Fig. 12.4 Typical shapes of maximum likelihood surfaces of the phylogenetic scaling factor lambda (λ). The *upper figures* show two examples, in which the surface has a distinct peak and only a narrow range of parameter values are likely. In contrast, the *bottom graphs* depict two cases in which the likelihood surface is rather flat, thus incurring a considerable uncertainty when choosing a single value. *Vertical red lines* give the values at the maximum likelihood. For the illustrative purposes, it is assumed that y-axes have the same scale

combined with multimodel inference for statistical weights (Fig. 12.5). We keep on emphasizing that our suggestions merely stand on theoretical grounds; the performance of model averaging in dealing with the uncertainty of model parameterizations awaits future tests (based on both empirical and simulated data).
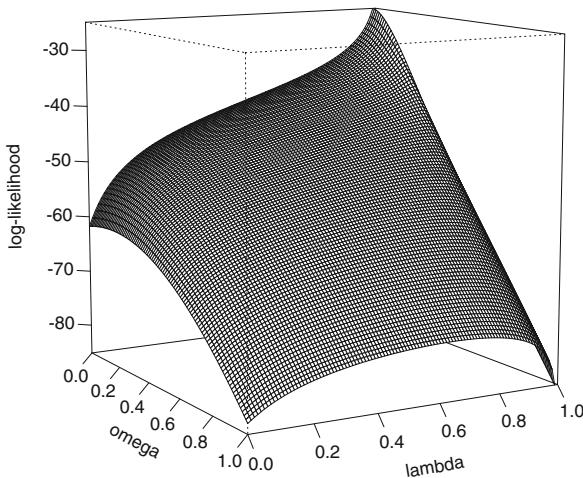
Fig. 12.5 Likelihood surface when the phylogenetic signal (lambda, λ) and the data heterogeneity (omega, ω) parameters are estimated in a set of models using different parameter combinations for the brain size/body size evolution in primates (data are shown in Fig. 12.1). The surface shows the log-likelihoods of a large number of fitted models that differ in their λ and ω parameters. These parameters are allowed to vary between 0 and 1 (with steps of 0.01) in all possible combinations. For a definition of ω, see Fig. 12.2

12.3.5   The Performance of Different Phylogenetic Comparative Methods

The logic of model selection can also be applied to assess whether any particular comparative method is more appropriate than others. For example, in a meta-analysis, Jhwueng (2013) estimated the goodness of fit of four phylogenetic comparative approaches. He collected more than a hundred comparative datasets from the published literature, to which he applied the following methods to estimate the phylogenetic correlation between two traits: the non-phylogenetic model (i.e., treating the raw species data as being independent), the independent contrasts method (Felsenstein 1985), the autocorrelation method (Cheverud et al. 1985), the PGLS method incorporating the Ornstein-Uhlenbeck process (Martins and Hansen 1997), and the phylogenetic mixed model (Hadfield and Nakagawa 2010; Lynch 1991). Model fits obtained for different approaches were compared based on AIC, which revealed that the non-phylogenetic model and the independent contrasts model offered the best fit. However, the parameter estimates for the phylogenetic correlation were quite similar across models, indicating that the studied comparative methods were generally robust to describe evolutionary patterns present in interspecific data.

12.4 Further Applications

So far, we mostly focused on the potential that the IT framework provides in association with the PGLS framework, when models are fitted with ML. However, multimodel inference also makes sense in a broader context, and related issues are known to exist in a range of other phylogenetic situations. We provide some examples below (without the intention of being exhaustive), but further applications can also be envisaged. This short list may illustrate that the benefits of multimodel inference can be efficiently exploited in relation to interspecific data.

A typical model selection problem is present in phylogenetics, when the interest is to find the best model that describes patterns of evolution for a given nucleotide or amino acid sequence. As briefly discussed in Chap. 2 (but see in-depth discussion in Alfaro and Huelsenbeck 2006; Arima and Tardella 2012; Posada and Buckley 2004; Ripplinger and Sullivan 2008), several models have been developed to deal with different substitution rates and base frequencies that ultimately influence the evolutionary outcome. The reliance on different models for phylogenetic reconstructions can result in phylogenetic trees that vary in their branching pattern and the underlying stochastic processes of nucleotide sequence changes that generate branch lengths. Given that a priori information about the appropriateness of different evolutionary models is generally lacking, those who wish to establish a phylogenetic hypothesis from molecular sequences are often confronted with a model selection problem. Accordingly, several evolutionary models need to be fitted to the sequence data, and the one that offers the best fit (e.g., as revealed by likelihood ratio test or an AIC-based comparison or Bayesian methods) should be used for further inferences about the phylogenetic relationships.

An intriguing example for the application of IT approaches in the phylogenetic context is the use of likelihood methods to detect temporal shifts in diversification rates. By fitting a set of rate-constant and rate-variable birth–death models to simulated phylogenetic data, Rabosky (2006) investigated which rate parameter combination (e.g., rate constant or rate varying over time) results in the model with the lowest AIC. The results suggested that selecting the best model in this way causes inflated Type I error, but when correcting for such error rates, the birth–death likelihood approach performed convincingly.

Eklöf et al. (2012) applied IT methods to understand the role of evolutionary history for shaping ecological interaction networks. The authors approached the effect of phylogeny by partitioning species into taxonomic units (e.g., from kingdom to genus) and then by investigating which partitioning best explained the species' interactions. This comparison was based on likelihood functions that described the probability that the considered partition structure reproduces the real data obtained for nine published food webs. Furthermore, they also used marginal likelihoods (i.e., Bayes factors) to accomplish model selection across taxonomic ranks. The major finding of the study was that models considering taxonomic

partitions (i.e., phylogenetic relationships) offered better fit to the data, and food webs are best explained by higher taxonomic units (kingdom to class). These results show that evolutionary history is important for understanding how community structures are assembled in nature.

Depraz et al. (2008) evaluated competing hypotheses about the postglacial recolonization history of the hairy land snail *Trochulus villosus* by using AIC-based model selection. They compared four refugia hypotheses (two refugia, three refugia, alpine refugia, and east–west refugia models) that could account for the phylogeographic history of 52 populations. The four hypotheses were translated into migration matrices, with maximum likelihood estimates of migration rates. These models were challenged with the data, and Akaike weights were used to make judgments about relative model support. This exercise revealed that the model considering the two refugia hypothesis overwhelmingly offered the best fit.

In a phylogenetic comparative study based on ancestral state reconstruction, Goldberg and Igić (2008) investigated 'Dollo's law' which states that complex traits cannot re-evolve in the same manner after loss. When using simulated data and an NHST approach (likelihood ratio tests), they found that in most of the cases the true hypothesis about the irreversibility of characters was falsely rejected. However, when using appropriate model selection (based on AIC-based IT methods), the false rejection rate of 'Dollo's law' was reduced.

Alfaro et al. (2009) developed an algorithm they called MEDUSA, which is an AIC-based stepwise approach that can detect multiple shifts in birth and death rates on an incompletely resolved phylogeny. This comparative method estimates rates of speciation and extinction by integrating information about the timing of splits along the backbone of a phylogenetic tree with known taxonomic richness. Diversification analyses are carried out by first finding the maximum likelihood for the per-lineage rates of speciation and extinction at a particular combination of phylogeny and species richness and then comparing these models across different combinations.

Further examples, e.g., for detecting convergent evolution based on stepwise AIC (Ingram and Mahler 2013) and for revealing phylogenetic paths based on the C-statistic Information Criterion (von Hardenberg and Gonzalez-Voyer 2013) can be found in Chaps. 18 and 8, respectively.

12.5   Concluding Remarks

What we have proposed here are several approaches to exploit the strengths of IT-based inference in the context of phylogenetic comparative methods. Using IT methods such as model selection in combination with phylogenetic comparative methods seems to offer the potential to elegantly solve problems which otherwise would be hard to tackle. Other IT methods such as model averaging allow dealing with phylogenetic uncertainty by explicitly incorporating it into the analysis and

exploring to what extend it compromises certainty about the results. Taken together, IT-based methods offer a great potential since they relieve researchers from the need of making arbitrary and/or poorly grounded decisions in favor of one or the other model. Instead, they allow dealing easily with such uncertainties or, at least, allow an assessment of their magnitude (among the set of potential models). Uncertainty is at the heart of our understanding about nature; thus, statistical methods are needed that appreciate this attribute instead of neglecting it.

We need to stress, though, that our propositions are based on theoretical grounds and need to be tested before they can be trusted. Particularly, simulation studies (e.g., along the design in Box 12.1) seem suitable for this purpose since they allow to investigate to what extend our propositions are able to reconstruct 'truth' which otherwise (i.e., in the case of using empirical data) is simply unknown. Simulation studies are warranted because the use of AIC (and other IT metrics) to non-nested models (which was largely the case here) is somewhat controversial (Schmidt and Makalic 2011). Another cautionary remark is that we refrained ourselves to suggest that only the IT-based model selection can be used to address the problems we raised. We envisage this discussion to serve as an initiative for comparative studies to consider the suggested methods as additions to the already existing toolbox, which yet await further exploitation.

Since the philosophy of IT-based inference is rather different from that of the classical NHST approach and since the two approaches are quite frequently mixed in an inappropriate way (e.g., selecting the best model using AIC and then testing it using NHST), we feel that some warnings on the use of the IT approach might be useful (particularly for those who were trained in NHST): IT-based inference does not reveal something like a 'significance,' and the two approaches must not be combined (Burnham and Anderson 2002; Mundry 2011). In the context of our propositions, this means that at least part of them naturally preclude the use of significance tests. This is particularly the case when sets of models with different combinations of predictors and/or sets of different phylogenetic trees are investigated. The end result of such an exercise is a number of AIC values associated with a set of models. Selecting the best model using AIC and then testing its significance is inappropriate. Rather, one could model average the estimates and their standard errors (but not the P values!) or also the fitted values and explore to what extend these vary across the different trees. Furthermore, one could use Akaike weights to infer about the relative importance of the different predictors. However, some of our proposed approaches might not necessarily and completely rule out the use of classical NHST. In fact, we do not argue against using NHST, which we regard as a scientifically sound approach if used and interpreted correctly. What we recommend is to not combine the use of significance tests with any of the approaches we suggested and draw inference solely on the basis of IT methods (e.g., Akaike weights or evidence ratios).

References

Alfaro ME, Huelsenbeck JP (2006) Comparative performance of Bayesian and AIC-based measures of phylogenetic model uncertainty. Syst Biol 55(1):89–96. doi:10.1080/10635150500433565

Alfaro ME, Santini F, Brock C, Alamillo H, Dornburg A, Rabosky DL, Carnevale G, Harmon LJ (2009) Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. Proc Natl Acad Sci. doi:10.1073/pnas.0811087106

Arima S, Tardella L (2012) Improved harmonic mean estimator for phylogenetic model evidence. J Comput Biol 19(4):418–438. doi:10.1089/cmb.2010.0139

Arnold C, Matthews LJ, Nunn CL (2010) The 10kTrees website: a new online resource for primate hylogeny. Evol Anthropol 19:114–118

Bennett PM, Harvey PH (1985) Brain size, development and metabolism in birds and mammals. J Zool 207:491–509

Blomberg S, Garland TJ, Ives AR (2003) Testing for phylogenetic signal in comparative data: behavioral traits are more laible. Evolution 57:717–745

Bolker B (2007) Ecological models and data in R. Princeton University Press, Princeton and Oxford

Burnham KP, Anderson DR (2002) Model selection and multimodel inference: a practical information-theoretic approach. Springer, New York

Burnham KP, Anderson DR, Huyvaert KP (2011) AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. Behav Ecol Sociobiol 65(1):23–35

Butler MA, King AA (2004) Phylogenetic comparative analysis: a modeling approach for adaptive evolution. Am Nat 164(6):683–695. doi:10.1086/426002

Chamberlin TC (1890) The method of multiple working hypotheses. Science 15:92–96

Cheverud JM, Dow MM, Leutenegger W (1985) The quantitative assessment of phylogenetic constraints in comparative analyses: sexual dimorphism of body weight among primates. Evolution 39:1335–1351

Claeskens C, Hjort NL (2008) Model selection and model averaging. Cambridge University Press, Cambridge

Cohen J (1994) The earth is round ($p < .05$). Am Psychol 49(12):997–1003

Collar DC, O'Meara BC, Wainwright PC, Near TJ (2009) Piscivory limits diversification of feeding morphology in centrarchid fishes. Evolution 63:1557–1573

Collar DC, Schulte JA, Losos JB (2011) Evolution of extreme body size disparity in monitor lizards (Varanus). Evolution 65(9):2664–2680. doi:10.1111/j.1558-5646.2011.01335.x

Congdon P (2003) Applied bayesian modelling. Wiley, Chichester

Congdon P (2006) Bayesian statistical modelling, 2nd edn. Wiley, Chichester

de Villemereuil P, Wells JA, Edwards RD, Blomberg SP (2012) Bayesian models for comparative analysis integrating phylogenetic uncertainty. BMC Evol Biol 12. doi:10.1186/1471-2148-12-102

Depraz A, Cordellier M, Hausser J, Pfenninger M (2008) Postglacial recolonization at a snail's pace (Trochulus villosus): confronting competing refugia hypotheses using model selection. Mol Ecol 17(10):2449–2462. doi:10.1111/j.1365-294X.2008.03760.x

Eklof A, Helmus MR, Moore M, Allesina S (2012) Relevance of evolutionary history for food web structure. Proc Roy Soc B-Biol Sci 279(1733):1588–1596. doi:10.1098/rspb.2011.2149

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

Freckleton RP, Harvey PH, Pagel M (2002) Phylogenetic analysis and comparative data: a test and review of evidence. Am Nat 160:712–726

Gamerman D, Lopes HF (2006) Markov chain Monte Carlo: stochastic simulation for Bayesian inference. CRC Press, Boca Raton, FL

Garamszegi LZ (2011) Information-theoretic approaches to statistical analysis in behavioural ecology: an introduction. Behav Ecol Sociobiol 65:1–11. doi:10.1007/s00265-010-1028-7

Garamszegi LZ, Møller AP (2007) Prevalence of avian influenza and host ecology. Proc R Soc B 274:2003–2012

Garamszegi LZ, Møller AP (2012) Untested assumptions about within-species sample size and missing data in interspecific studies. Behav Ecol Sociobiol 66:1363–1373

Garamszegi LZ, Møller AP, Erritzøe J (2002) Coevolving avian eye size and brain size in relation to prey capture and nocturnality. Proc R Soc B 269:961–967

Goldberg EE, Igic B (2008) On phylogenetic tests of irreversible evolution. Evolution 62(11):2727–2741. doi:10.1111/j.1558-5646.2008.00505.x

Hadfield JD, Nakagawa S (2010) General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. J Evol Biol 23:494–508

Hansen TF (1997) Stabilizing selection and the comparative analysis of adaptation. Evolution 51:1341–1351

Hansen TF, Bartoszek K (2012) Interpreting the evolutionary regression: the interplay between observational and biological errors in phylogenetic comparative studies. Syst Biol 61:413–425

Harmon LJ, Losos JB, Jonathan Davies T, Gillespie RG, Gittleman JL, Bryan Jennings W, Kozak KH, McPeek MA, Moreno-Roark F, Near TJ, Purvis A, Ricklefs RE, Schluter D, Schulte Ii JA, Seehausen O, Sidlauskas BL, Torres-Carvajal O, Weir JT, Mooers AØ (2010) Early bursts of body size and shape evolution are rare in comparative data. Evolution 64(8):2385–2396. doi:10.1111/j.1558-5646.2010.01025.x

Hegyi G, Garamszegi LZ (2011) Using information theory as a substitute for stepwise regression in ecology and behavior. Behav Ecol Sociobiol 65:69–76. doi:10.1007/s00265-010-1036-7

Hunt G (2006) Fitting and comparing models of phyletic evolution: random walks and beyond. Paleobiology 32(4):578–601. doi:10.1666/05070.1

Hutcheon JM, Kirsch JW, Garland TJ (2002) A comparative analysis of brain size in relation to foraging ecology and phylogeny in the chiroptera. Brain Behav Evol 60:165–180

Ingram T, Mahler DL (2013) SURFACE: detecting convergent evolution from comparative data by fitting Ornstein-Uhlenbeck models with stepwise Akaike information criterion. Methods Ecol Evol 4(5):416–425. doi:10.1111/2041-210x.12034

Ives AR, Midford PE, Garland T (2007) Within-species variation and measurement error in phylogenetic comparative methods. Syst Biol 56(2):252–270

Iwaniuk AN, Dean KM, Nelson JE (2004) Interspecific allometry of the brain and brain regions in parrots (Psittaciformes): comparisons with other birds and primates. Brain Behav Evol 30:40–59

Jhwueng D-C (2013) Assessing the goodness of fit of phylogenetic comparative methods: a meta-analysis and simulation study. PLoS ONE 8(6):e67001. doi:10.1371/journal.pone.0067001

Johnson JB, Omland KS (2004) Model selection in ecology and evolution. Trends Ecol Evol 19(2):101–108

Konishi S, Kitagawa G (2008) Information criteria and statistical modeling. Springer, New York

Kutsukake N, Innan H (2013) Simulation-based likelihood approach for evolutionary models of phenotypic traits on phylogeny. Evolution 67(2):355–367

Lajeunesse MJ (2009) Meta-analysis and the comparative phylogenetic method. Am Nat 174(3):369–381. doi:10.1086/603628

Legendre P, Lapointe FJ, Casgrain P (1994) Modeling brain evolution from behavior: a permutational regression approach. Evolution 48(5):1487–1499. doi:10.2307/2410243

Link WA, Barker RJ (2006) Model weights and the foundations of multimodel inference. Ecology 87:2626–2635

Lynch M (1991) Methods for the analysis of comparative data in evolutionary biology. Evolution 45(5):1065–1080

Martins EP (1996) Conducting phylogenetic comparative analyses when phylogeny is not known. Evolution 50:12–22

Martins EP, Hansen TF (1997) Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. Am Nat 149:646–667

Massart P (ed) (2007) Concentration inequalities and model selection: ecole d'eté de probabilités de Saint-Flour XXXIII - 2003. Springer, Berlin

Mundry R (2011) Issues in information theory-based statistical inference–a commentary from a frequentist's perspective. Behav Ecol Sociobiol 65(1):57–68

Mundry R, Nunn CL (2008) Stepwise model fitting and statistical inference: turning noise into signal pollution. Am Nat 173:119–123

Nakagawa S, Hauber ME (2011) Great challenges with few subjects: Statistical strategies for neuroscientists. Neurosci Biobehav Rev 35(3):462–473

O'Meara BC, Ané C, Sanderson MJ, Wainwright PC (2006) Testing for different rates of continuous trait evolution using likelihood. Evolution 60(5):922–933. doi:10.1111/j.0014-3820.2006.tb01171.x

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

Pagel M, Meade A, Barker D (2004) Bayesian estimation of ancestral character states on phylogenies. Syst Biol 53(5):673–684

Pagel M, Meade A (2006) Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. Am Nat 167(6):808–825

Posada D, Buckley TR (2004) Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst Biol 53(5):793–808. doi:10.1080/10635150490522304

R Development Core Team (2013) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.orgS

Rabosky DL (2006) Likelihood methods for detecting temporal shifts in diversification rates. Evolution 60(6):1152–1164

Ripplinger J, Sullivan J (2008) Does choice in model selection affect maximum likelihood analysis? Syst Biol 57(1):76–85. doi:10.1080/10635150801898920

Scales JA, King AA, Butler MA (2009) Running for your life or running for your dinner: what drives fiber-type evolution in lizard locomotor muscles? Am Nat 173:543–553

Schmidt D, Makalic E (2011) The behaviour of the Akaike information criterion when applied to non-nested sequences of models. In: Li J (ed) AI 2010: advances in artificial intelligence, vol 6464. Lecture Notes in Computer Science. Springer, Heidelberg, pp 223–232. doi:10.1007/978-3-642-17432-2_23

Stephens PA, Buskirk SW, Hayward GD, Del Rio CM (2005) Information theory and hypothesis testing: a call for pluralism. J Appl Ecol 42(1):4–12

Symonds MRE, Moussalli A (2011) A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. Behav Ecol Sociobiol 65(1):13–21

Terribile LC, Olalla-Tarraga MA, Diniz JAF, Rodriguez MA (2009) Ecological and evolutionary components of body size: geographic variation of venomous snakes at the global scale. Biol J Linn Soc 98(1):94–109

Thomas GH, Freckleton RP, Székely T (2006) Comparative analyses of the influence of developmental mode on phenotypic diversification rates in shorebirds. Proc Roy Soc B-Biol Sci 273(1594):1619–1624

von Hardenberg A, Gonzalez-Voyer A (2013) Disentangling evolutionary cause-effect relationships with phylogenetic confirmatory path analysis. Evolution 67(2):378–387. doi:10.1111/j.1558-5646.2012.01790.x

Whitmee S, Orme CDL (2013) Predicting dispersal distance in mammals: a trait-based approach. J Anim Ecol 82(1):211–221. doi:10.1111/j.1365-2656.2012.02030.x

Whittingham MJ, Stephens PA, Bradbury RB, Freckleton RP (2006) Why do we still use stepwise modelling in ecology and behaviour? J Anim Ecol 75:1182–1189

Part III
Specific Models for Studying Evolutionary Mechanisms

Chapter 13
Simulation of Phylogenetic Data

Emmanuel Paradis

Abstract Simulating phylogenetic data is a powerful tool for evolutionists, but this can be a complicated task. This chapter gives an overview on the methods to simulate species traits, particularly on a phylogeny. We show that building from three fundamental models (Brownian motion (BM), Ornstein–Uhlenbeck (OU), and Markov chains (MC)), many biologically relevant scenarios can be simulated. We also review briefly the simulation of phylogenies and the available software for phylogenetic data simulation (PDS). The online materials give several examples, including some complex cases, using R.

13.1 Introduction

Biological evolution proceeds over various scales of time, space, and complexity. It is thus an ideal subject for computer simulation. Phylogenetic data simulation (PDS) can be fascinating because of its power, but also intimidating because of its sophistication, which may disconnect it from real biological problems. The goal of this chapter is to show how recent developments in PDS can be incorporated in the general framework of the phylogenetic comparative method (PCM) and help to answer some fundamental biological questions.

13.2 Two Examples

We start with two examples to illustrate how simulations can help in evolutionary studies. The first one considers the phylogeny of salamanders of the genus *Plethodon* as reconstructed by Highton and Larson (1979) and shown in Fig. 13.1.

E. Paradis (✉)
Institut de Recherche pour le Développement, Montpellier, France
e-mail: emmanuel.paradis@ird.fr
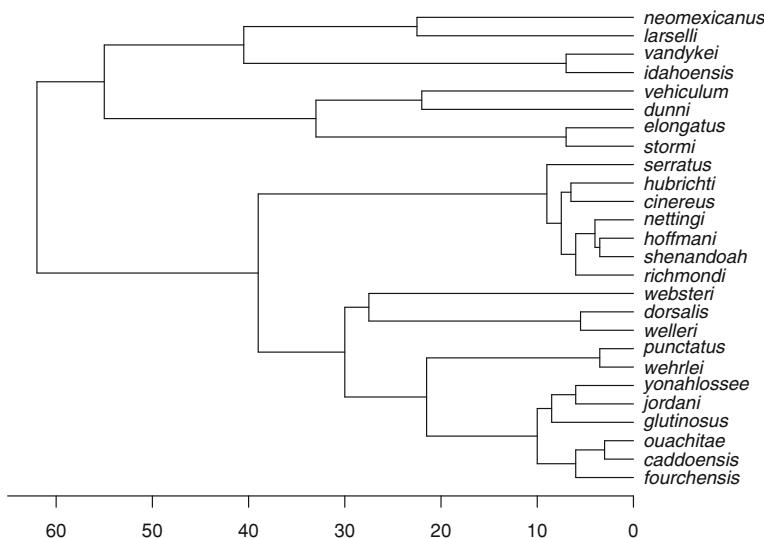
Fig. 13.1 Phylogeny of 26 species of the genus *Plethodon*

These amphibians display several interesting biological features such as having no lung. Kozak et al. (2009) studied body size evolution of this group by considering four categories of size. One way to investigate this question is to simulate a discrete trait along the phylogeny of these salamanders and to look at the patterns generated under different parameter values. For simplicity, we consider here two categories of body size and assume that change is equally likely in both directions. The parameter to be varied here is the rate of change 'large body' $\rightleftarrows$ 'small body'. We can simulate the evolution of body size many times, starting arbitrarily with 'large' as the root state, and examine the frequencies of both states for the tips of the tree. Figure 13.2 shows the proportion of large-bodied species for three values of the rate of change. This shows clearly that, if this rate is moderate or high (0.1 or higher), we should expect both body size categories to be in approximately equal proportions. Of course, this conclusion holds for any trait or variable evolving along this phylogeny under the same model. The phylogeny is also a parameter of the simulation, and our conclusion may not hold for a different one—this can be checked by running the simulations with a different tree.

This simple example shows that simulations can be used to make theoretical predictions without the complicated mathematical developments typically used by theoreticians. Furthermore, this illustrates a fundamental point about prediction under a probabilistic model of evolution: the frequency of species in a particular state is a random variable, and simulations give a good picture of its distribution (a result that would be very difficult to obtain with theoretical equations of this model).
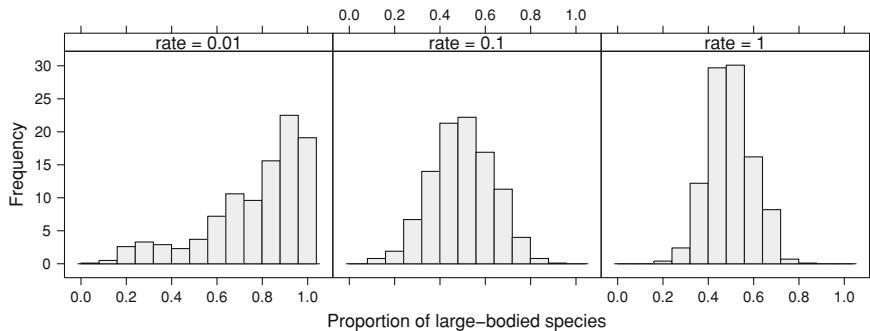
Fig. 13.2 Distribution of the proportion of large-bodied species over 1,000 simulations starting with a large-body root state and evolving along the tree shown on Fig. 13.1

The second example has a more statistical motivation: it shows how some statistical results about PCMs that may be difficult to conceptualize can be illustrated easily with simulations (see Chap. 1). We consider the case of two traits that do not evolve along the phylogeny because, for instance, they are subject to selection so that closely related species are not likely to be more similar than distant ones. We want to know the consequences of using Felsenstein's (1985) method of phylogenetically independent contrasts (PIC) to analyse these data. The online materials show an example of R code to perform this task. By repeating simulations, a large number of times when the null hypothesis is true (no correlation between the two traits), we can estimate the type I error rate of the PIC method that is the probability to reject the null hypothesis when it is true. Using 10,000 replications, we found that the type I error rates were 0.119, 0.171 and 0. 191, for sample sizes of 10, 50 and 100 species, respectively. This result shows two things. First, in this situation, the probability of rejecting the null hypothesis when it is true is substantially higher than 0.05. So, taking into account phylogenetic relationships when this should not be done can lead to wrong conclusions. Second, this probability is higher for increasing sample size. This is again a result that is not trivial and would be extremely difficult to demonstrate with formal mathematical developments.

13.3 Traits

Simulating traits on a phylogeny requires a model of evolution. Classically, two models are used for continuous traits: Brownian motion (BM) and Ornstein–Uhlenbeck (OU). For discrete traits, the basic model is a Markov chain (MC). These three models can be simulated in either continuous or discrete time (Fig. 13.3 and Box 13.1).

Box 13.2 gives some details on the structure of these three models. It appears that evolution of a single trait can be simulated using four basic parameters: σ^2 (the
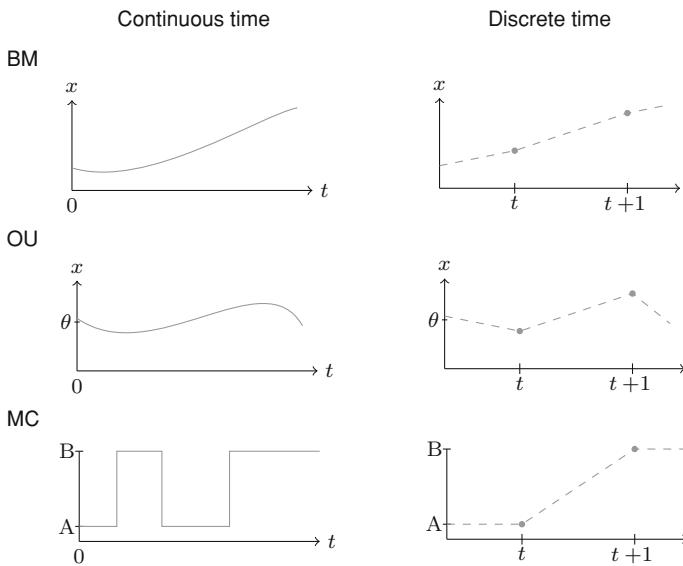
Fig. 13.3 The three fundamental models of trait evolution simulation in continuous or discrete time. *BM* Brownian motion; *OU* Ornstein–Uhlenbeck; *MC* Markov chain

rate of evolution in the BM and OU models), θ (the optimum value of the traits in the OU model), α (the strength of selection toward θ), and r (the rate(s) of change among states in MC models). The simplest models of trait simulation are thus defined with one (or three) constant parameter(s). However, from a biological point of view, it is relevant to be able to vary these parameters in order to simulate data under biologically relevant scenarios. It is important here to distinguish two layers of complication of these models.

The first layer considers a single trait and how its parameter(s) of evolution may vary, suppose we want these parameters to vary among the different branches of a tree. Typically, a PDS will run successively along the edges of the tree starting from the root. Continuous-time equations can be used with the branch length as t, and the value of the trait at a node is passed to the daughter lineages. Therefore, there is no difficulty in making the parameter values different among the branches.

Box 13.1 Mathematics of PDS

Deterministic Versus *Stochastic Simulations.*
Classically, two kinds of simulations are distinguished: deterministic (involving the application of successive equations) and stochastic (involving random numbers). Some simulations of evolution may be deterministic (for instance, a model of infinite population with selection), but in the vast majority of cases, PDS will be stochastic. Figure 13.4 shows an example of the contrast between the two types of simulation using an OU model. Each

simulation starts from $x = 0$ and 'moves' progressively toward the optimum value $\theta = 10$. This shows that considering stochasticity in the model depends on the parameter values, of course, but also on the dynamics under way: stochasticity is less important when x is increasing because the dynamics here is dominated by the 'attraction' toward θ.

Continuous- Versus *Discrete-Time Simulations.*

Time is continuous but the operations on a computer are discrete because they are treated sequentially by the processor. So considering time as continuous or discrete in PDS makes sense depending on the context of the study. Conceptually, a discrete-time model is simpler and may be written in a generic way as $x_{t+1} = f(x_t)$, where x is the simulated variable, and f is a function. A continuous-time model is more complicated to derive and often starts from a differential equation that specifies how the simulated variable changes during a very short time interval $\mathrm{d}t$:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(x_t) \ .$$

There are two possible ways to use such a differential equation in a simulation. The first way is to find an analytical solution that will have typically a form $x_t = f(x_0, t)$, where x_0 denotes the initial state of x. The second way is to solve the differential equation numerically, often because it is too complicated to find an analytical solution; in that case, the task is similar to a discrete-time simulation. Note that 'time' is denoted as t in both cases but they have very different meanings: in continuous time, t is a continuous variable, whereas in discrete time, t, $t + 1, \ldots$, are successive time steps (see Figs. 13.3 and 13.5).

The choice of considering time as continuous or discrete in a simulation depends on the situation. An important consideration is whether an analytical equation is available with t continuous. Another important consideration is whether time (in absolute units, e.g. years) is a crucial ingredient of the model. There are a number of analytical formulae that can be used in PDS, some of them are mentioned in this chapter or can be found in the online materials.

Suppose now that we want to make the parameters dependent on absolute time: their values will be defined by a function of time. For instance, consider the BM parameter and assume that it follows a function denoted as $s(t)$. Continuous-time formulations become more complicated as they require integration over time (e.g. Molini et al. 2011): if the integral of s is easy to calculate, then there is no
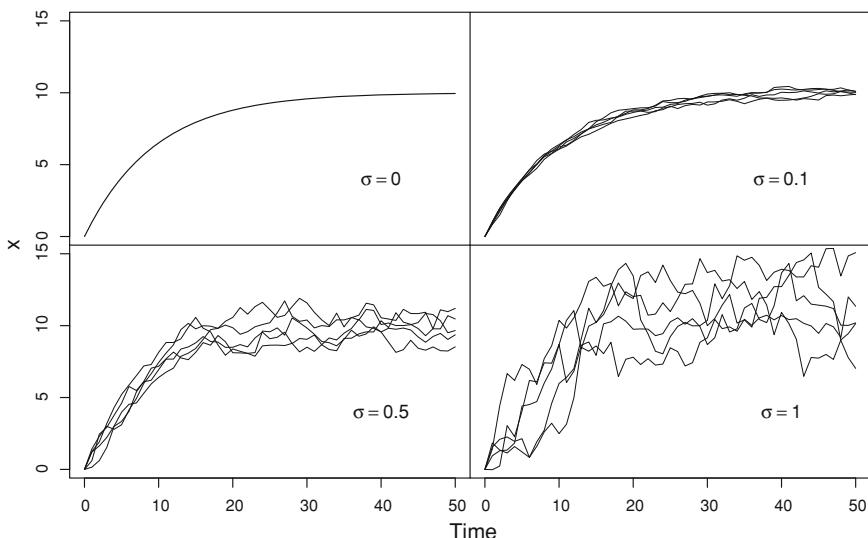
Fig. 13.4 How stochasticity changes the dynamics of a simulation. The values of x were simulated with a discrete-time OU model with $\theta = 10$, $\alpha = 0.1$, $x_0 = 0$, and σ as indicated on each graph (five replications were done). See how stochasticity is more obvious when $x \approx \theta$

computational difficulty. If not, such integrals can be evaluated numerically. An alternative is to consider discrete-time models in which case σ^2 varies at each time step (Fig. 13.5). This approach is attractive, especially since continuous-time OU or MC models are difficult to analyse (see Massey and Whitt 1998). It follows that a wide range of biologically interesting scenarios can be simulated as a time-dependent formulation can be applied to several forms of variation (which can be considered separately or in combination):

- Intrinsic: the parameter values depend on species variables, (body mass, metabolic rate, etc.) which themselves change through time.
- Extrinsic: the parameter values are related to global variables (climate) for all species (or a subset).
- Random: the parameters vary randomly either completely or partially, possibly with some form or temporal auto-correlation.

The second layer of complication is when considering several traits simultaneously. They may evolve independently (involving no additional parameter), or they may be linked, which will usually be modeled with a covariance function or a functional relationship. Depending on the number of traits and on the nature of their links, this may imply one or several additional parameters. Some examples are given in the online materials http://www.mpcm-evolution.com/.

(a)

(b)

(c)
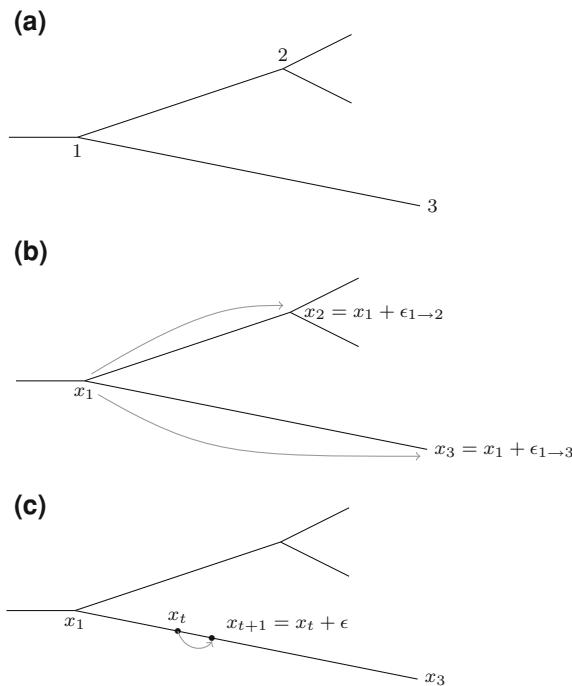
Fig. 13.5 Simulating a trait x on a tree under a Brownian motion model with continuous or discrete time. **a** A portion of a phylogenetic tree showing three nodes numbered *1* to *3*. **b** Simulation in continuous time: the values of x for the nodes *2* and *3* are generated directly from their ancestor. The ε's are random variates, following a normal distribution with mean zero and variance given by the product of the parameter σ^2 with the corresponding branch lengths (hence the subscripts). **c** Simulation in discrete time: the branch lengths are broken down in elementary time steps, and the trait values are generated by successive applications of the equation, which is shown here for a single step. The ε's are now random variates, following a normal distribution with mean zero and variance equal to the rate parameter σ^2 (t is not involved). It is clear that the continuous-time version is faster to compute, but the discrete-time one handles easily external variables, which may change independently of the tree

Box 13.2 Basic Models of Single Trait Evolution

The *Brownian motion* (BM) model has many applications such as modeling the random walk of particles or the evolution of continuous traits under random drift. The basic operation of simulating under a BM model is to add to the value of the trait at time t, x_t, a random normal variate with mean zero and variance σ^2. This parameter is the only parameter of the BM model and is often referred to as the rate of evolution in the biological literature. Clearly, the larger the value of σ^2 the faster x will change. Because change occurs in any direction, the expected value of x is equal to the initial value x_0

while the variance around this expectation is given by $\sigma^2 t$ so that the probability to observe $x_t = x_0$ decreases when t increases. The covariance between two tips (i.e., the similarity between the values of x measured on two species) is given by the product of σ^2 with the distance between the root of the tree and the most common recent ancestor of both tips (which measures the quantity of shared evolution between them).

In the *Ornstein–Uhlenbeck* (OU) model, change in x is 'attracted' toward a value denoted as θ with a 'strength' controlled by the parameter α. So there are two additional parameters compared to the BM model. In discrete time, setting $\alpha = 0$ reduces the OU model to the BM model, but in continuous time, this equivalency works only in the limit for very small value of α (when α tends to zero) because of a division by 2α (Gillespie 1996). The covariance between two tips of a tree under this model is given by $\sigma^2 \exp(-\alpha \mathrm{d})/2\alpha$, where d is the distance between both tips measured on the tree; so, by contrast to the BM model, this does not depend on shared evolution.

Markov chains (MC) offer a very general framework to model the evolution of discrete traits. Figure 13.3 shows the simplest MC model with two states (denoted as A and B) and a single parameter so that changes in both directions A → B and B → A occur at the same speed. Here, the parameterization differs between the continuous- and discrete-time cases. In continuous time, the parameter is a rate r with the constraint $0 \le r$, which quantifies the fastness of change during a very short time interval (so short that multiple changes are impossible). The rate matrix Q is built with the rates, and its diagonal elements are set so that its rows sum to zero. If a transition between two states is not directly possible, zero is entered in the corresponding entry of Q. The probabilities of change among the states of x during a time interval t are calculated with the matrix exponential of the product Qt. The result of this operation is a matrix denoted as P and its rows sum to one. In discrete time, this framework does not apply and the parameter is a probability of change p ($0 \le p \le 1$) arranged in a matrix P whose rows sum to one.

13.4 Trees

The simulation of trees is a vast topic in computer science (Siltaneva and Mäkinen 2002). We shall limit ourselves to the methods and algorithms relevant for PCMs.

13.4.1 Speciation–Extinction Trees

Simulating trees of various shapes (i.e., focusing on their topology) have received a lot of interest during the 1980s and 1990s; see the thorough review by Mooers and Heard (1997). An issue that has received a lot of attention is whether different models are more likely to generate unbalanced topologies than balanced ones (Page 1991). Over the last decade, speciation–extinction trees (also known as birth–death trees) have become increasingly popular as they are more biologically realistic since they consider the process of successive random speciation and extinction events over time. The parameters are the speciation and extinction rates (or probabilities) denoted as λ and μ.

Box 13.3 gives the rationale behind most algorithms for simulating trees, as well as drawing some connections with simulation of traits. The methods are described below in the section on software.

13.4.2 Non-ultrametric Trees

Non-ultrametric trees are often considered when one wants to simulate trees estimated from molecular sequence data. The rationale behind this method is the fact that branch lengths in phylogenies reconstructed from molecular data are assumed to be the product of time and substitution rate (Felsenstein 2004). In that case, the procedure is to first simulate an ultrametric tree (typically, a birth–death one) and then transform its branch lengths by multiplying them by the rates of molecular substitution (Brown and Yang 2011). Unless this last parameter is constant throughout the tree, the resulting tree will be non-ultrametric.

13.4.3 Trees and Traits Jointly

The joint simulation of traits and trees is a good illustration of how different approaches can be used and combined. In Paradis (2005), a discrete-time approach was used to simulate the evolution of a continuous trait that affects speciation probability. FitzJohn (2012) used a continuous-time approach to simulate the evolution of a discrete trait that affects speciation and extinction rates: he used a time-to-event formulation where the events are speciation, extinction, and transitions among states. The following section gives more details on this model and its variants.

13.5 Software

Software for PDS can be grouped in two categories: integrated and stand-alone. Integrated software is part of a set of programs or code devoted to phylogenetics or evolutionary biology. Stand-alone software has programs that do not need additional programs to run on a computer. Integrated software for phylogenetics has become increasingly popular over the last decade, and three computer languages are now widely used: Perl, Python, and R. However, only the last one offers tools to simulate traits under almost all models we have seen above. Therefore, we shall focus on R in this section and mention briefly the tools available in the two other languages.

Box 13.3 Time-to-Event and PDS

When simulating trees, there is a duality between continuous- and discrete-time methods. In continuous time, the speciation (λ) and extinction (μ) parameters are effectively rates and may be larger than one. Kendall (1948) studied the behaviour of a general model, where λ and μ are functions of time using differential equations. He derived probability formulae for the distribution of population size (which could be individuals or species). More recently, Maddison et al. (2007), FitzJohn et al. (2009), and Etienne and Rosindell (2012) used a similar approach for other models of diversification. Explicit formulae are different compared to the simulation of traits because we consider here discrete events (births and deaths). It is interesting to look at some of its details because this shows how different mathematical tools appear as different views of the same problem and complement each others. Suppose an individual (or a species) is exposed to death (or extinction) with rate μ. We can model this situation with a Markov chain since there are two discrete states with the rate matrix Q given by:

$$\begin{array}{c} \\ \text{alive} \\ \text{dead} \end{array} \begin{array}{cc} \text{alive} & \text{dead} \\ \begin{bmatrix} -\mu & \mu \\ 0 & 0 \end{bmatrix} \end{array}$$

We remind that the diagonal elements are set so that the rows sum to zero. Because there is only one possible transition (alive $\rightarrow$ dead), the matrix exponential of Qt can be written directly:

$$\begin{array}{c} \\ \text{alive} \\ \text{dead} \end{array} \begin{array}{cc} \text{alive} & \text{dead} \\ \begin{bmatrix} e^{-\mu t} & 1 - e^{-\mu t} \\ 0 & 1 \end{bmatrix} \end{array}$$

The top-right term of this matrix is the cumulative density function (CDF) of the exponential distribution. (The parameter of the exponential distribution

is usually denoted as λ, which we use here to denote the speciation parameter, or sometimes as h for hazard rate.) Indeed, the probability of dying at time t is also the probability of surviving until but not after t. Thus, we find the well-known result that if μ is constant, the survival times follow an exponential distribution with density $\mu e^{-\mu t}$ (i.e., the derivative of the CDF). If μ varies through time, an integration over t is done similar to what we have seen in the main text for traits, so the probability of surviving until at least time t is:

$$\exp\left[-\int_0^t \mu(u)\mathrm{d}u\right].$$

The event considered may not be only extinction but also speciation, so that these equations can be used to simulate the branch lengths and the topology of a tree. As above, time dependence in λ and μ can be included either with explicit formulae if their primitives can be found, or using numerical integration (Paradis 2011; Hallinan 2012).

The interesting thing about writing down the relationship between an MC model and a time-to-event formulation is that it appears how to handle more complex models (see Sect. 4.3). For instance, Etienne and Rosindell (2012) considered a model where species are either 'good' or 'incipient' each with their own parameters (particularly, they assumed $\lambda = 0$ for the incipient species) and used differential equations to simulate phylogenies from it and derive some formulae.

In the discrete-time setting, $\lambda(t)$ and $\mu(t)$ are probabilities and what has been seen above about traits can also be applied to simulation of trees.

13.5.1 Perl and Python

The Bio::Phylo module (Vos et al. 2011) has several functions to simulate trees under some simple models: birth–death and coalescent models. The simulation is stopped when a specified number of tips are reached.

The package DendroPy (Sukumaran and Holder 2010) can simulate trees in discrete or continuous time. The parameters λ and μ can be constant or vary randomly at the node of the tree according to a normal distribution with variance specified by the user. More complex models can be simulated with 'multiple stages' or by extending an existing tree.

The package Cass provides several functions to generate branching times under various models of speciation–extinction and different sampling schemes (e.g. Stadler 2009).

13.5.2 R

By contrast to Perl or Python, authors of R packages have adopted common data structures to code phylogenetic trees and traits, so inter-operability among those packages is enhanced. Consequently, a greater variety of tools are available under this environment. Currently, there are *ca.* 100 packages available for phylogenetics with R, though only a few provide tools for PDS. It is difficult to know precisely "what does what" among these packages because most of them are modified through regular releases. So, the list of functions and tools below is susceptible to change over time. To keep track of these changes, a list of packages and their main functionalities is maintained on the Comprehensive R Archive Network (CRAN) as a 'Task View for phylogenetics'.[1] The online materials provide a number of examples of PDS with R.

Ape (Paradis et al. 2004) provides three functions to simulate any kind of trait(s) under any model (Table 13.1). All these functions require a tree as input. For instance, the command `rTraitCont(tr)` will simulate a single trait under a BM model with the rate parameter $\sigma = 0.1$ along the tree named tr. Several options of these functions control the model, the parameter values, or whether to output the values of the trait for the nodes of the tree (they are always output for the tips). These functions allow us to simulate trait(s) under all models described above, especially `rTraitMult`, which can handle several traits (Table 13.2).

Diversitree (FitzJohn 2012) offers a variety of functions to perform joint simulation of traits and trees (Table 13.3). The basic model is called 'binary state speciation and extinction' (BiSSE) introduced by Maddison et al. (2007). This model considers a binary trait taking the value 0 or 1 evolving under an asymmetric Markov model (so the changes $0 \rightarrow 1$ and $1 \rightarrow 0$ occur at different rates). The phylogeny itself evolves with speciation and extinction parameters that depend on the state of the binary trait. Therefore, this model has six parameters. This basic model has been extended in several directions, such as the 'MuSSE' model where the trait has more than two states, or the 'QuaSSE' model where the trait is quantitative (Tables 13.3 and 13.4). This package includes also several utility functions to extract information from the output of the simulations (e.g. `history.from.sim.discrete`). This helps to visualize the tree together with the transitions of the character(s).

Ape includes some basic functions to simulate trees (Table 13.5). The function `rtree` generates a tree by successive random splits; the branch lengths are

---

[1] http://cran.r-project.org/web/views/Phylogenetics.html

Table 13.1 Functions to simulate traits on a phylogeny with ape

Function	What it simulates	Comments
rTraitCont	Single continuous trait with BM, OU or custom model	Fast code for BM and OU models
		Parameters can be branch specific
		Custom model can be anything
rTraitDisc	Single discrete trait with Markov models	Any number of states
		Total control on the model
		Rates can be branch specific
rTraitMult	Multivariate models	Total control on the model
		Any number of traits

Table 13.2 Some options of the functions described in Table 13.1 (*C* continuous, *D* discrete, *M* multivariate)

Option	Value	Type of trait(s)
phy	The tree	CDM
model	The model	CDM
ancestor	Logical value	CDM
root.value	Value of the trait at the root	CDM
sigma	σ	C
alpha	α	C (model = 'OU')
theta	θ	C (model = 'OU')
k	The number of states	D
rates	r	D
states	Labels for the states	D
p	The number of traits	M

Table 13.3 Functions to simulate traits and trees with diversitree

Function	What it simulates	Comments
tree.bd	Tree with λ and μ constant	
tree.yule	Id. with $\mu = 0$	
tree.bisse	Tree and a binary trait	Uses the BiSSE model
tree.bisseness	Id	A variant of the BiSSE model
tree.musse	Tree and a discrete trait	Uses the MuSSE model
tree.musse.multitrait	Tree and several discrete traits	Id
tree.quasse	Tree and a continuous trait	Uses the QuaSSE model
tree.classe	Tree and a discrete trait	Uses the cladogenetic SSE model
tree.geosse	Tree and a geographic range	Uses the GeoSSE model

Table 13.4 Options of the functions described in Table 13.3. The simulation is stopped when either 'max.taxa' or 'max.t' is reached

Option	Value
pars	Parameters (numerical values)
max.taxa	Maximum number of species 'alive'
max.t	Maximum time span of the tree
include.extinct	Logical value
x0	Initial value of the trait

Table 13.5 Functions to simulate trees with ape (n number of tips)

Function	What it simulates	Comments
rtree	Non-ultrametric tree	n fixed
rcoal	Ultrametric tree	Assumes Θ constant; n fixed
rbdtree	Ultrametric tree	Fossil lineages not output
		Simulation ends after fixed time; n variable
rlineage	Non-ultrametric tree	All lineages output
		Simulation ends after fixed time; n variable

simulated independently (by default from a uniform distribution that can be changed by the user). `rcoal` simulates a coalescent tree with constant population parameter Θ. Coalescent trees are mostly relevant for population questions, but this function is a quick way to generate an ultrametric tree with a fixed number of tips. Two other functions use any model of temporal change in λ and μ to simulate birth–death trees in continuous time (as described in Paradis 2011): `rbdtree`, which outputs an ultrametric tree with only the lineages surviving until the end of the simulation, and `rlineage`, which outputs a tree with all lineages. Both functions simulate trees on a fixed time.

Another important class of algorithms to generate birth–death trees is based on conditioning on a fixed value of n. The package `TreeSim` provides a function, `sim.bd.taxa`, to simulate phylogenies with a fixed number of tips, whereas `sim.bd.taxa.age` simulates with fixed number of tips and fixed time. It can also simulate incomplete phylogenies under various sampling schemes (Stadler 2009, 2011).

13.5.3 Stand-alone Programs

A vast number of programs offer the possibility to simulate trees and/or characters, most of them with limited possibilities (e.g. simulating a tree with constant speciation and extinction rates, or a trait under a BM model). We limit ourselves to a few that are interesting in the present context.

- EREM can simulate a random tree and binary (0/1) characters with a random model that pertains to the evolution of genomes (mainly insertions and deletions). The parameters can vary randomly along the tree. It is implemented in Matlab (a C++ version was announced a few years ago). http://carmelab.huji.ac.il/software/EREM/erem.html.
- EvolSimulator can perform complex simulations of genes and genomes, including heterogeneous mutation regimes among lineages, variable patterns of selective pressure across sequences, paralogy, and horizontal gene transfer. It is implemented in C++. http://acb.qfab.org/acb/evolsim/.
- Phylocom simulates phylogenies in discrete time together with up to five traits with a pseudo-Brownian motion model. Ancestral competition can be simulated too. It is implemented in C and OS-specific binaries are available for download. http://phylodiversity.net/phylocom/.

References

Brown RP, Yang Z (2011) Rate variation and estimation of divergence times using strict and relaxed clocks. BMC Evol Biol 11:271

Etienne RS, Rosindell J (2012) Prolonging the past counteracts the pull of the present: protracted speciation can explain observed slowdowns in diversification. Syst Biol 61:204–213

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

Felsenstein J (2004) Inferring phylogenies. Sinauer Associates, Sunderland, MA

FitzJohn RG (2012) Diversitree: comparative phylogenetic analyses of diversification in R. Meth Ecol Evol 3:1084–1092

FitzJohn RG, Maddison WP, Otto SP (2009) Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. Syst Biol 58:595–611

Gillespie DT (1996) Exact numerical simulation of the Ornstein-Uhlenbeck process and its integral. Phys Rev E 54:2084–2091

Hallinan N (2012) The generalized time variable reconstructed birth–death process. J Theor Biol 300:265–276

Highton R, Larson A (1979) The genetic relationships of the salamanders of the genus *Plethodon*. Syst Zool 28:579–599

Kendall DG (1948) On the generalized "birth-and-death" process. Ann Math Stat 19:1–15

Kozak KH, Mendyk RW, Wiens JJ (2009) Can parallel diversification occur in sympatry? Repeated patterns of body-size evolution in coexisting clades of North American salamanders. Evolution 63:1769–1784

Maddison WP, Midford PE, Otto SP (2007) Estimating a binary character's effect on speciation and extinction. Syst Biol 56:701–710

Massey WA, Whitt W (1998) Uniform acceleration expansions for Markov chains with time-varying rates. Ann Appl Prob 8:1130–1155

Molini A, Talkner P, Katul GG, Porporato A (2011) First passage time statistics of Brownian motion with purely time dependent drift and diffusion. Phys A 390:1841–1852

Mooers AØ, Heard SB (1997) Inferring evolutionary process from phylogenetic tree shape. Quart Rev Biol 72:31–54

Page RDM (1991) Random dendrograms and null hypotheses in cladistic biogeography. Syst Zool 40:54–62

Paradis E (2005) Statistical analysis of diversification with species traits. Evolution 59:1–12

Paradis E (2011) Time-dependent speciation and extinction from phylogenies: a least squares approach. Evolution 65:661–672

Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290

Siltaneva J, Mäkinen E (2002) A comparison of random binary tree generators. Comput J 45:653–660

Stadler T (2009) On incomplete sampling under birth-death models and connections to the sampling-based coalescent. J Theor Biol 261:58–66

Stadler T (2011) Simulating trees with a fixed number of extant species. Syst Biol 60:676–684

Sukumaran J, Holder MT (2010) DendroPy: a Python library for phylogenetic computing. Bioinformatics 26:1569–1571

Vos RA, Caravas J, Hartmann K, Jensen MA, Miller C (2011) BIO: Phylo-phyloinformatic analysis using perl. BMC Bioinformatics 12:63

Chapter 14
Use and Misuse of Comparative Methods in the Study of Adaptation

Thomas F. Hansen

Abstract The comparative method can be used to test hypotheses of adaptation by comparing groups of species that meet different adaptive challenges. This requires attention to phylogenetic correlations and to historical lags in achieving adaptation. The modern phylogenetic comparative method has provided some partial solutions to these problems, but the field has also suffered from a systemic lack of demand for biological justifications of its statistical procedures. Consequently, assumptions have been made for statistical convenience and are often inconsistent with the relevant biology. I argue that common comparative tests of adaptation, including Brownian-motion based phylogenetic linear models and inferred-changes methods based on reconstructing ancestral states, violate essential characteristics of adaptation as a biological process. I discuss the requirements for biologically consistent comparative analysis of adaptation, and I review work toward this goal.

14.1 Introduction

Ever since Darwin the comparative method has been a major tool for studying adaptation on macroevolutionary timescales. By comparing species living in different niches or environments, we can look for systematic differences in biological traits and relate these to functional needs in the environment. Consider deer antlers. The males of most species in the deer family (Cervidae) sport antlers that are used in sexual displays. These are cast after the mating season and regrown each year. This pattern suggests an influence of sexual selection. We can construct a comparative test for the involvement of sexual selection by comparing deer

T. F. Hansen (✉)
Department of Biology, CEES, University of Oslo, Blindern,
P.O. Box 1066, 0316 Oslo, Norway
e-mail: Thomas.Hansen@bio.uio.no

species likely to experience different strengths of sexual selection. This can be based on ecological data on the severity of competition among males for females. For example, one could compare the size of antlers in species with monogynous versus polygynous mating systems. In a now classical paper, Clutton-Brock et al. (1980) did this by comparing the size of antlers across species with different-sized breeding groups, and they indeed found larger antlers in species with more competitive mating systems.

There are, however, many difficulties with such analyses. Perhaps the most fundamental of these relates to the fact that we are working with observational data and not with controlled experiments. This means that there are usually many possible interpretations of any one pattern, and many unknown or at least uncontrolled factors that can influence the result. Hence, it is crucial that comparative tests of adaptation are not done in isolation from other biological information. Studies have to start from specific hypotheses with a priori biological motivation, and the truism that hypotheses can only be falsified and not confirmed is even more to the point in observational than it is in experimental studies. In the case of antlers, we have a priori reasons to evoke sexual selection, but there are many different sexual-selection hypotheses to choose among, and even if some form of sexual selection is operating, it is almost certainly not the only relevant selective force. The growth of antlers requires considerable resources, and different diets may put different costs on antler growth (Geist 1998). For example, Geist argued that one factor that allowed the gigantic antlers of the Irish elk was a mineral-rich diet of willow (see also Gould 1998). Antlers may also be a hindrance of movement and the costs of this may depend on the habitat (open or closed), the predation pressure, the shape of the antler, and the size of the animal. Antlers are also integrated with the rest of the organism. Strong static and ontogenetic allometries exist within species, and the positive evolutionary (among-species) allometry (Fig. 14.1) that culminates in the relatively huge antlers in large-bodied species such as the Irish elk could be a non-adaptive side effect of body-size evolution (Gould 1973, 1974, 1977). Sexual selection may also act on alternative traits, and because of the positive allometry, large antlers in deer with competitive mating systems could be a non-adaptive side effect of selection on sexual size dimorphism. The mating system correlates with body size, and Clutton-Brock et al. (1980) even suggested that the positive evolutionary allometry of antler size may be a result of adaptive evolution of larger male bodies in more polygynous mating systems.

Adaptation is not instantaneous, and the ancestry of the species must be considered when evaluating its antler traits. For example, although the Irish elk had relatively huge antlers that may have deviated positively from the evolutionary allometry, so do the antlers of its closest living relatives, the fallow deers (Fig. 14.1). Lister et al. (2005) and Hughes et al. (2006) obtained ancient DNA from the Irish elk and showed that it indeed was related to fallow deers in the genus *Dama* (Fig. 14.2). Hence, the antlers of the Irish elk may reflect ancestral constraints and cannot automatically be assumed to be fully adapted to the "current" environment of the species.
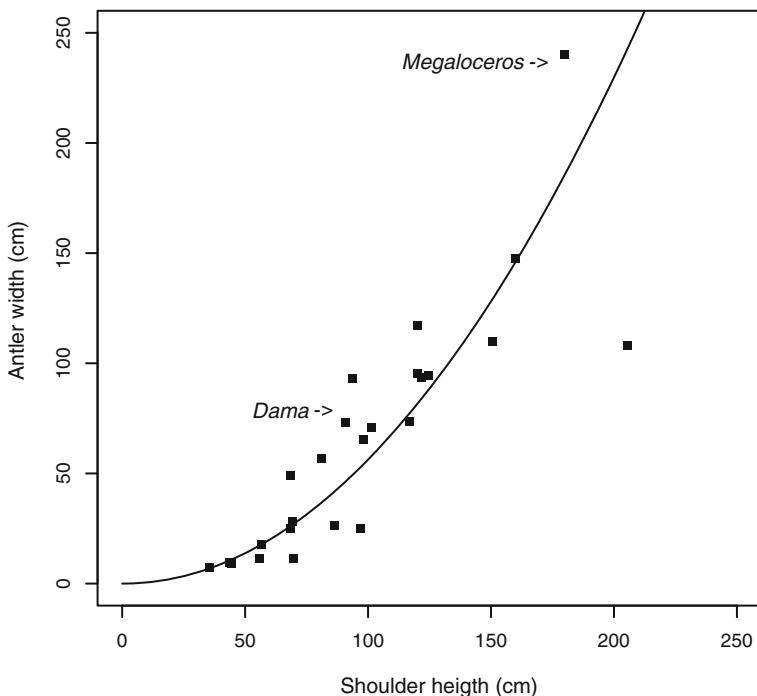
Fig. 14.1 Allometry of antler size against shoulder height in deer based on data from Clutton-Brock et al. (1980) with the Irish elk added (data from Gould 1974). Allometric curve estimated from non-phylogenetic least-squares regression on log-scale. The Irish elk (*Megaloceros*) and its closest living relative, the fallow deer (*Dama*), both deviate positively from the allometric relationship

Such ancestral constraints generate phylogenetic covariance (=phylogenetic correlation) between species and may thus violate the independence assumptions of standard statistical methods. This problem came in focus in the 1970s and contributed to a negative view of species comparison as non-rigorous methods that could only suggest, but not test, hypotheses. The emergence of so-called phylogenetic comparative methods that dealt with this statistical problem was therefore an important development in evolutionary biology. The foundational paper of Felsenstein (1985) presenting the method of independent contrasts showed how statistically correct analyses could be conducted based on a phylogeny with branch lengths and an assumed model of evolution. Together with rapid growth in the quality and availability of molecular phylogenetic information, this led to a renaissance of large-scale comparative studies and in the last decade, we have seen proliferation of ever more sophisticated methods for comparative analysis (see Martins 2000; Garland et al. 2005; Cooper et al. 2010; Freckleton et al. 2011; Nunn 2011; Stone et al. 2011; O'Meara 2012; Pennell and Harmon 2013; and this volume for some twenty-first century reviews).
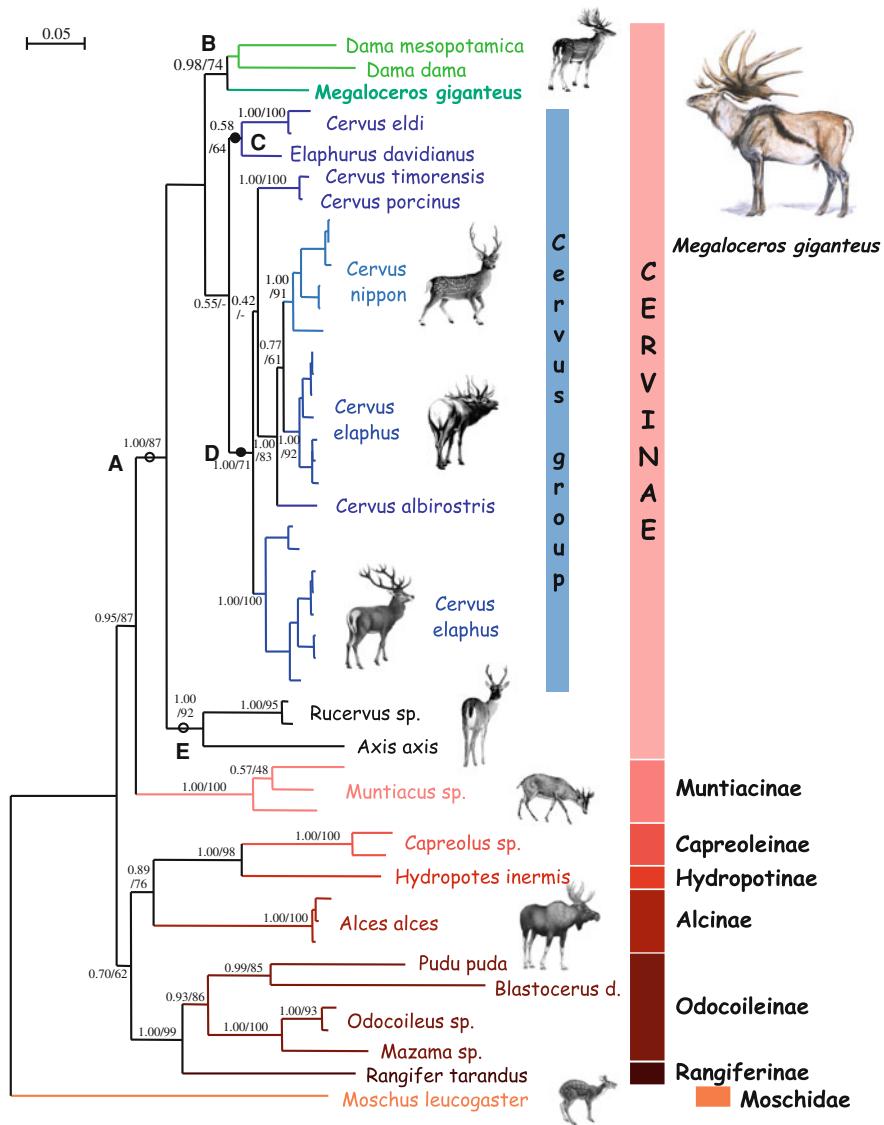
Fig. 14.2 Molecular phylogeny for the deer family including the Irish elk based on ancient DNA. Reprinted from Hughes et al. (2006) with permission from Elsevier

Developers of phylogenetic comparative methods have been concerned with solving statistical problems and have sometimes lost sight of the biological contexts in which the methods are to be used. Nowhere is this more evident than in the study of adaptation. For a long time, phylogenetic comparative methods for quantitative traits were built almost exclusively around a single model of evolutionary change, the Brownian-motion model, which introduces serious interpretational problems

when it is used to represent adaptive evolution. It is plainly inconsistent with adaptation toward an optimal state in given niche (Hansen and Orzack 2005). Starting with Hansen (1997) and Butler and King (2004), alternative methods that were more appropriate for testing adaptation were proposed, but have only recently been in regular use. Over the last few years, these methods have seen rapid development. Here, I review some of these developments and discuss the logic of and problems with the comparative study of adaptation.

14.2 The Concept of Adaptation

To better understand how the comparative method can be informative about adaptation, it is useful to clarify how I think about the concept and to outline some of the different notions of adaptation that have played a role in the comparative methods literature.

While it is possible to talk loosely about adaptation of a whole organism to a niche, as in saying that the polar bear is adapted to the arctic, quantitative studies must focus on specific biological traits. Adaptations are also for something specific. A trait is an adaptation for some task, X, if it helps perform X. The polar bear's white fur is an adaptation for camouflage against the arctic snow if it does help make the bear less visible and this improves the bear's fitness. Note that we may speak of adaptation in this sense even if the relevant task disappears. The white fur is an adaptation for camouflage against the snow even if the snow disappears and the species come to exist in a snow-free environment, say the zoo.

While all (genetic) adaptation is due to selection, not all selection leads to adaptation. A trait may be under selection for a number of reasons, but only selection causally generated by the success in performing task X will build adaptation for task X. This observation is related to Sober's (1984) distinction between selection *for* a trait and selection *of* a trait. A trait Y may be selected because it is correlated with some other trait under selection. This is called indirect selection, and while it is a potent mechanism for evolution, it is no more a mechanism for generating adaptations in Y than non-selective factors such as genetic drift or gene flow. Furthermore, a trait may be selected for more than one reason, and selection of trait Y for task Z will not generate adaptation for task X except by circumstance.

It seems reasonable to expect most traits to be influenced by several sources of both direct and indirect selection. The antlers of a deer may be under direct selection for impressing females, but this must be seen against a background of indirect selection acting on body size and mineral metabolism, and it will have to compete and interact with direct selection originating in male–male conflict and predation.

For this reason, I think adaptation is best understood in the context of a balance of forces. In fields such as behavioral ecology and life-history theory, investigations often start with the assumption that a trait is optimized by selection, and

adaptation for task X is studied by asking how variations in the need for X may shift the position of the optimum (Michell and Valone 1990; Reeve and Sherman 2001). This assumes that there are constrains from other selective forces. By saying that the shape of the antlers of a deer is adapted to attract females, we mean that female preferences have been able to shift the optimal antler shape away from the value it would have had if the females did not care. Note that this makes adaptation into a question of degree. The antlers may not be, and are indeed not expected to be, optimal for attracting females. It is enough that the female preference has some influence on the fitness function of the antlers.

In Hansen (1997), I proposed to understand the comparative study of adaptation in a similar way. As is done in within-species optimality studies of adaptation, we can start by assuming that the traits are optimal and then use the comparative method to compare the position of the optima in different environments. A comparative study of deer with monogynous versus polygynous mating systems can reveal whether antlers tend to evolve toward different states in the two situations. If they do, this is evidence that the antler optima are influenced by mating system or some variables correlated with mating system.

This view of the comparative study of adaptation is consistent with the most common ways of studying adaptation within species, but it differs dramatically from how adaptation is usually conceived in phylogenetic or historical studies. For example, there is a cladistic tradition for doing comparative studies of adaptation based on identifying trait changes and associating these with states or changes in the environment (e.g., Ridley 1983; Coddington 1988; Baum and Larson 1991; Brooks and McLennan 1991; Maddison 1994; Larson and Losos 1996). In this approach, only apomorphisms are candidate adaptations. This view is also reflected in Gould and Vrba's (1982) distinction between adaptation and exaptation, where the term adaptation is restricted to traits that have demonstrably originated due to the adaptive function (also favored by Sober 1984). In contrast, the broad non-historical concepts of adaptation favored by many, and in practice used by all, neontologists (Reeve and Sherman 1993) differ by focusing on whether traits are maintained by selection for the "adaptive" (including exaptive) function regardless of how the trait originated.

In comparative studies, these different notions of adaptation may lead to dramatically different approaches and conclusions. While the historical approach is limited to identifiable changes (or even apomorphisms), the non-historical approach will also utilize trait maintenance, absence of change, as evidence of adaptation (e.g., Williams 1992; Frumhoff and Reeve 1994; Westoby et al. 1995; Hansen 1997; Price 1997; Reeve and Sherman 2001). To exaggerate, a cladist may hold that the relatively huge antlers of the Irish elk cannot be adaptations to anything in the elk's environment, because the trait is ancestral (i.e., shared with the fallow deers). The counter argument is that the large-sized antlers must be adaptations for something because they have been maintained for long periods of time in balance with obvious selection pressures to reduce them. In Hansen (1997), I argued that we have comparative evidence for hypsodonty, high-crowned cheek teeth, being an adaptation to grazing in horses, because the trait has been

maintained in all grazing horses and it has not evolved in browsing horses. A cladist may instead say we have little evidence as a parsimony analysis reveals only a single shift to hypsodonty within this group and this is not sufficient statistical evidence to draw any conclusions.

Ironically, Gould did not heed his own rallying cry of "stasis is data" when discussing adaptation.

14.3   The Logical Structure of the Comparative Study of Adaptation

Sober (2008) points out that comparative studies involve a shift in explanandum from why a trait has a certain value to why there is a difference between groups in this value. That is, we shift from trying to explain why the antlers of Irish elk have an average span of 3 m to explain why polygynous and monogynous deer have different antler sizes. I want to argue, however, that this shift in explanandum does not need to entail a shift in explanans. In both cases, the explanation can be based on the assumption that the trait is at or near an optimum and the explanatory goal is to identify the factors that determine the positions or differences of optima.

Hansen and Bartoszek (2012) formalized this as follows. Let Y be the state of the trait we are studying, say antler size, and let us assume that this trait is at or close to a fitness optimum determined as a function of a number of factors, X, as $Y = f(X_1, \ldots, X_m)$. Assuming we know the function, f, and the exact state of all relevant factors, we could make a perfect prediction of the position of the optimum. In practice, however, we can only know, or at least measure, one or a few relevant factors, and we do not know the exact functional relationship between these and the optimal state. It is still possible to test hypotheses of adaptation. Let us say we want to test whether antler size is influenced by sexual selection, and we have observations of the strength of sexual selection from several species. This could come from direct measures of variance in mating success or from data on mating systems and some theory linking mating system to strength of sexual selection (e.g., arguing that sexual selection on males is stronger in polygynous mating systems). Let X_1 be our measure of strength of sexual selection, and rewrite the model as

$$Y = b_0 + b_1 X_1 + r(X_1, \ldots, X_m), \tag{14.1}$$

where $r(X_1, \ldots, X_m) = f(X_1, \ldots, X_m) - (b_0 + b_1 X_1)$ are biologically determined residual deviations from the model. Obviously, these residuals will be different in different species due to different values of the X-variables. It is still possible to test the influence of our focal variable X_1 by estimating the coefficient b_1. This will give the linear influence of X_1 on the optimum. For example, if X_1 is taken to be an indicator of polygynous versus monogynous mating systems, b_1 will be the average difference in optimal antler size for the two mating systems. I want again to underscore that this does not assume that there is one unique optimum for each

mating system. Each species has its own optimum, $f(X_1, \ldots, X_m)$, determined by many factors. The comparative method works by identifying systematic effects of some focal variables above the background noise due to changes in other unmeasured factors and then testing whether these effects are consistent with theoretical predictions. If not, the theory should be revised.

Many issues with comparative methods revolve around the model residuals, which may violate any of the standard linear model assumptions. Because the residuals depend on biological factors that may be shared among related species, they cannot be assumed to be independent. The residuals may also depend on the focal factor or on variables that are related with the focal factor. Given that some factors may also be discrete or of major influence, the residuals are not necessarily normally distributed. While these problems are serious, they can all be diagnosed and at least partially dealt with. The focus of the modern phylogenetic comparative method has been to solve the non-independence problem. We will now look at how this is done.

14.4 Phylogenetic Comparative Methods

Most comparative analyses are standard regression or ANOVA types of analyses with species trait means as the dependent variable. In this setting, the phylogenetic comparative method is a standard linear model modified to account for phylogenetic correlations in the residuals. Formally, the model can be written as

$$\mathbf{y} = \mathbf{D}\beta + \mathbf{r} \tag{14.2}$$

$$\mathbf{r} \sim N(0, \mathbf{V}) \tag{14.2a}$$

where $\mathbf{y}$ is a vector of species observations, $\mathbf{D}$ is a design matrix with predictor variables, $\boldsymbol{\beta}$ is a vector of parameters to be estimated, and $\mathbf{r}$ is a vector of residuals assumed to be normally distributed with mean zero and, a not necessarily diagonal, variance matrix, $\mathbf{V}$. If the $\mathbf{D}$ and $\mathbf{V}$ are specified, generalized least squares (GLS) estimates can be used to obtain unbiased minimum-variance estimates of the parameter vector $\boldsymbol{\beta}$, and if $\mathbf{D}$ and $\mathbf{V}$ depend on additional parameters with unknown values, as they typically do in adaptation models, then so-called estimated GLS, where the additional parameter values are estimated by maximum likelihood, can be used (e.g., Martins and Hansen 1997; Chaps. 5 and 6).

While this is straightforward in principle and can also be extended to generalized linear models such as logistic regression (Martins and Hansen 1997; Hadfield and Nakagawa 2010; Ives and Garland 2010; Chap. 9), the crux is how to model the $\mathbf{D}$ and $\mathbf{V}$ matrices. The standard approach, which forms the basis of nearly all phylogenetic comparative methods, is to leave the design matrix fixed and model the covariances in $\mathbf{V}$ as proportional to shared branch lengths on a phylogeny. This is justified if the response and predictor variables jointly follow a multivariate Brownian-motion process (Felsenstein 1985). In this case, regressions

of variables on each other are linear and residuals are normally distributed with covariances proportional to shared branch lengths. This is the basis for the method of independent contrasts, which may be regarded as an algorithm for implementing GLS when residual covariances are proportional to shared branch lengths.

It is essential to realize, however, that other evolutionary processes lead to different patterns. Hansen and Martins (1996) give an overview of the phylogenetic correlation patterns expected under different evolutionary processes, and many do not provide a simple proportionality of covariance to branch lengths (see Chap. 15). Furthermore, it is also typical that model residuals have patterns of phylogenetic signal that are different from the patterns in the data themselves (Labra et al. 2009).

14.5 Problems with the Standard Phylogenetic Comparative Method

Confusing phylogenetic signal in the data (e.g., the traits) with phylogenetic signal in model residuals is perhaps the most common error in the application of modern phylogenetic comparative methods (Fig. 14.3). The core assumptions of independence, normality, and homoscedascity that are made in the standard regression and ANOVA models apply to the residuals from the model and not to the response or predictor variables per se. Still, it is entirely common to see tests for phylogenetic signal being conducted on the data and used to justify the use, or not, of phylogenetic corrections. This is a fundamental error and has no doubt lead to many misapplications of phylogenetic corrections (see Hansen and Orzack 2005; Labra et al. 2009; Revell 2010; Hansen and Bartoszek 2012 for discussions of the problem).

Figure 14.3a illustrates a common pattern in cross-species regression analyses. This is essentially the standard textbook illustration of the need for phylogenetic corrections, but this pattern is in fact totally consistent with the assumptions of standard non-phylogenetic regression. Even if there is a strong phylogenetic signal in both the response (y-axis) and the predictor variable (x-axis), there is no indication of a phylogenetic pattern in the residual deviations. This situation arises when adaptation is rapid. Then, we do not expect a phylogenetic signal in model residuals, but if related species tend to occur in similar environments (i.e., having similar values of their predictor variables), then we still expect a phylogenetic signal in the response variable. Correcting for phylogeny in this situation is throwing the baby out with the bathwater.

My impression from the relatively few studies that have reported phylogenetic signal in model residuals is that this situation is more common than situations with strong phylogenetic signal in model residuals as illustrated in Fig. 14.3b. If so, then the application of phylogenetic comparative methods has done more harm than good in the study of adaptation. Standard non-phylogenetic methods would usually have been a better choice than methods based on independent contrasts and the like.
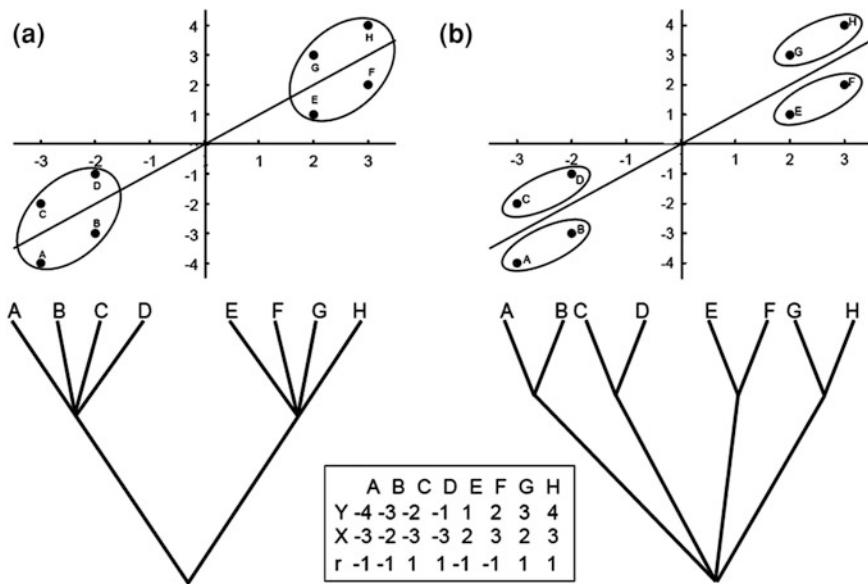
Fig. 14.3 Sources of phylogenetic effects. The trait Y on the y-axes is adapting to an environment X on the x-axes, but species A–H deviate from the regression line. In **a**, the species are divided into two clades as shown in the phylogeny, and even if both Y and X are associated with the phylogeny (all species in a clade have the same sign of Y and X values), the residual deviations, r, from the line are not associated with the phylogeny (species within a clade are deviating in different directions). In **b**, the species within a clade are deviating in similar directions, and a phylogenetic GLS would be appropriate. In **a**, the phylogenetic effect in the trait is generated by the association of the trait with a phylogenetically-structured environment, while in **b**, it is also influenced by phylogenetically correlated maladaptations. Developed in collaboration with A. Labra and modified from Labra et al. (2009)

Furthermore, this problem is but one instantiation of a suite of problems deriving from a statistically convenient, but biologically unjustified, separation of "non-phylogenetic" adaptation from "phylogenetic" residual deviations. For example, methods such as phylogenetic autocorrelation (e.g., Cheverud et al. 1985) and phylogenetic eigenvector regression (e.g., Diniz-Filho et al. 1998) are based on explaining trait variation with one or more descriptors of phylogenetic distance (e.g., phylogenetic eigenvectors) and analyzing trait–trait or trait–environment relations in the reminder. This removes phylogenetically structured adaptation from consideration. While phylogenetic (and spatial) eigenvector regressions are useful descriptive tools, they are not well suited for estimating adaptation (see Freckleton et al. 2011; Diniz-Filho et al. 2012 for debate).

Figure 14.4 illustrates how any adaptive process that is sufficiently slow to generate a phylogenetic signal in model residuals will also generate systematic deviations from the optimal state that would manifest as a lag or "bias" in the mean structure of the model (Hansen et al. 2008; Hansen and Bartoszek 2012;
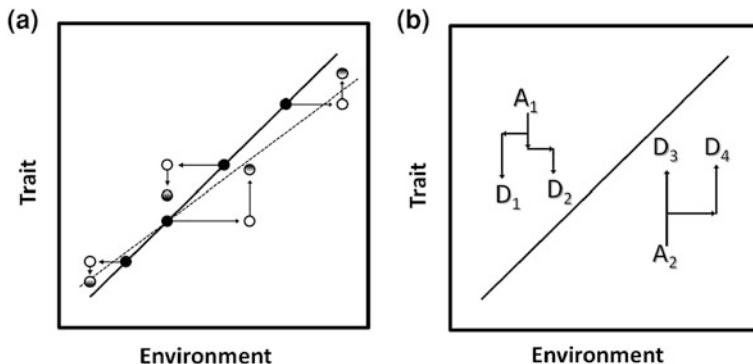
Fig. 14.4 Maladaptation, optimality, and phylogenetic effects. **a** How maladaptation flattens the evolutionary regression. A set of ancestral species (*black dots*) sit on a line describing the (primary) optimal relation between the trait and the focal environment. Secondary environmental changes may shift the species off the line in random directions (*horizontal arrows* to *open dots*), leading to a more shallow "evolutionary regression" (*dotted line*). The species then undergo adaptive evolution (*vertical arrows*), and the evolutionary regression will approach the optimal regression as the ancestral maladaptation is reduced. **b** How maladaptation leads to residual correlation. Two maladapted ancestral species, "*A*", speciate while evolving toward their (primary) optimal states. The descendant species, "*D*", deviate from optimality in a correlated manner. This correlation disappears as they reach their optima. Inspired by Sober (2008), and developed in collaboration with A. Labra

Bartoszek et al. 2012). This shows that the past history of the predictor variables should enter into the design matrix, **D**, and the weighting of the past should match with the degree of phylogenetic signal in the residual variance matrix, **V**. If not, estimates of the optimal relation to the predictor variables will be biased. For example, estimated regression slopes may be more shallow than the optimal relation between variables as illustrated in Fig. 14.4a. This means that the standard practice of modeling residual variance and mean structure separately is inconsistent with evolution toward optimal states. Species simply cannot provide unbiased information about optimal trait–environment relations at the same time as they retain ancestral residual correlations with other species (Fig. 14.4b).

Until the mid 2000s, almost all phylogenetic comparative analyses of continuous traits were based explicitly or implicitly on the assumption that model residuals evolve as an undirected Brownian motion. The Brownian motion has the property that the expected state of a descendant must equal the state of its ancestor. Hence, there is no mechanism for a systematic decrease of ancestral discrepancies between trait and environment. Instead, the degree of maladaptation will increase gradually through time due to the undirected random changes. Hansen and Orzack (2005) pointed out that this leads to what they called the problem of inherited maladaptation. If the tip species live in different environments, then the environment must have changed somewhere on the phylogeny, and then the change in the predictor variable implies that the species must jump to a new state where it has the same residual deviation as it had before. It inherits the maladaptation of its
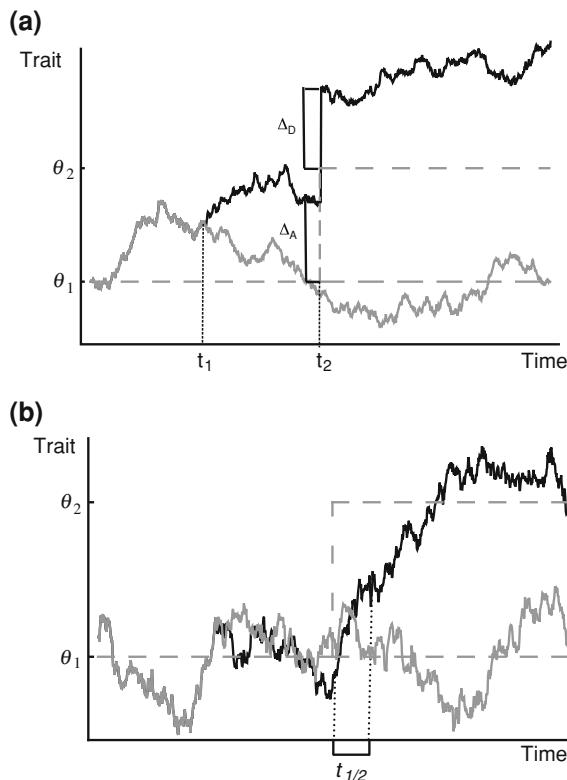
Fig. 14.5 Stochastic-process models of adaptation along a (two-species) phylogeny. **a** Evolution follows a Brownian motion. The ancestral species splits into two (*black* and *gray*) at time t_1, and at time t_2, one of the sister species (the *black*) experience an environmental change as its optimum (*dashed gray lines*) changes from θ_1 to θ_2. In a standard phylogenetic comparative analysis, this would be modeled by adding the difference $(\theta_2 - \theta_1)$ to the species (as the model is the sum of the environmentally predicted value and residuals evolving as Brownian motion). Note how this implies that ancestral maladaptation (Δ_A) must be transferred to maladaptation in the new environment (Δ_D) even if the ancestral species happened to be close to the new optimum. Note also how the Brownian motions tend to drift further and further away from the optimum. Maladaptation is increasing (on average) even in a constant environment. **b** Same as **a** except that the species now evolve according to an Ornstein–Uhlenbeck process around their optima. Here, the species tend to be pulled toward their optima if they diverge too far, and note how the two species lose their ancestral correlation much faster than with the Brownian motion. When the environment changes, the (*black*) species is gradually pulled toward its new optimum. The phylogenetic half-life ($t_{1/2} = \ln 2/\alpha$) is the time it takes to get halfway there (on average)

ancestor even when this implies it must jump across or away from its "optimal" state (Fig. 14.5a). This shows that, while the standard model can be used to estimate the statistical influence of an environmental variable on a trait, it is not a biologically coherent model of adaptation to fixed optima, and its estimates should not be taken as estimates of optimal states.

14.6 Modeling Adaptation

A way to approach the above problems is to base the comparative method on the Ornstein–Uhlenbeck process in place of the Brownian motion (Fig. 14.5; and see Chap. 15). The Ornstein–Uhlenbeck process models trait change as a sum of a white noise and a deterministic pull toward a particular state (the "optimum"). The model is

$$\mathrm{d}y = -\alpha(y - \theta)\mathrm{d}t + \sigma\,\mathrm{d}W, \tag{14.3}$$

where $\mathrm{d}y$ is the trait change in a time interval $\mathrm{d}t$, θ is the "optimum," and the $\mathrm{d}W$ is independent normally distributed random variables with mean zero and variance proportional to $\mathrm{d}t$ (i.e., a white noise). The parameters α and σ describe, respectively, the strength of the pull toward the optimum and the standard deviation of the stochastic changes. If α is zero, the pull to the optimum disappears and the Brownian motion is regained.

In Hansen (1997), I proposed a solution to the above problems based on assuming that the predictor variables acted on the optimum, θ, and thus only indirectly on the trait. Hence, the focus of estimation was shifted from the direct relationship between trait and environment and on to the relationship between the environment and the optimal state. In this way, species are allowed to be influenced by past environments, to lag behind their current optimal state, and to deviate in manners consistent with the deviation of their relatives. In the simplest cases, the model predicts that past environments have an influence proportional to an exponential function of the elapsed time since the environment occurred multiplied by α and that the covariance between related species is proportional to the exponential of the time separating them multiplied by α. The exact equations for this and various extensions of the model can be found in Hansen (1997) and Hansen et al. (2008).

This setup aligns the comparative method with within-species optimality studies of adaptation. In both cases, adaptation is studied as the influence of variables on an assumed optimum. As I argued above, the main reason why species trait means are different is because they evolve around different local optima determined by many factors. Optimality studies are concerned with testing the effects of one or a few variables on the optimum, and residual deviations are mainly due to the fact that we cannot know all relevant variables. This means that residual deviations are not due to maladaptation in the general sense, but reflect "maladaptation" relative to a focal selective agent under study. To make this distinction explicit, I introduced the concept of a *primary optimum*, defined as the average optimum reached by a number of species evolving in the same niche for sufficient time to allow ancestral constraints to disappear (Hansen 1997). This term was inspired by Simpson's (1944) concept of a primary adaptive zone. The idea is that the niche or adaptive zone is described by one or a few factors, say the grazing adaptive zone being defined by using grasses as major food source. The hypothesis that a trait, say hypsodont teeth, is an adaptation to this niche would then predict

that the primary optimum for tooth crown height of grazers is different from the primary optimum of other dietary niches. The local optimum of particular species in this niche may deviate from the primary optimum due to differences in secondary variables that do not enter the niche definition, say body size, skull shape, or the availability of alternative food sources, but if the trait is significantly influenced by the primary niche, then we predict systematic differences between the primary optima.

More specifically, our approach starts as a linear model of the primary optimum, determined by one or a few predictor variables that will differ across species or parts of the phylogeny. It then uses an Ornstein–Uhlenbeck model for evolution around this, not necessarily constant, primary optimum. The simplest setup is to map different niches onto a phylogeny and then estimate the value of the primary optimum in each of these niches. If θ_j is the value of the primary optimum in niche j, Hansen (1997) showed that the predicted trait value for species i is

$$\hat{y}_i = c_{0i}y_{\mathrm{a}} + c_{1i}\theta_1 + c_{2i}\theta_2 + \cdots + c_{ki}\theta_k, \tag{14.4}$$

where y_{a} is the ancestral state at the root of the phylogeny, and the coefficients c_{ji} represent the influence of environmental state j on species i. Each period in the species' past history that was associated with a niche contributes a term $e^{-\alpha t_{\mathrm{e}}} - e^{-\alpha t_{\mathrm{b}}}$ to the coefficient of that niche with t_{e} and t_{b} being the times back to the end and the beginning of the period. Hence, each coefficient c_{ij} is the sum of such terms for each period in which the species i was associated with niche j. The coefficient c_{oi} equals $e^{-\alpha t_{\mathrm{r}}}$, where t_{r} is the time back to the root of the phylogeny. Hence, all coefficients are between zero and one, and they sum to one. The coefficient corresponding to a particular environmental state will be large when the species has spent a lot of its history associated with this state, but more recent associations are weighted more heavily than more ancient ones. The larger the rate of adaptation, α, the more the weighting is shifted toward recent environments, and when α approaches infinity, then only the current environment is weighted and the model converges on a standard non-phylogenetic linear model.

There are many variations and extensions of this basic model. Butler and King (2004) developed a method for evaluating and comparing different niche arrangements with information-theoretical criteria. See also Ingram and Mahler (2013) for further extensions (Chap. 18). This is useful because the Achilles heel of the method is the need for reconstructing the niches on a phylogeny. The test of adaptation relies on the reconstruction of the historical past, and as I will discuss below, ancestral-state reconstruction is often unreliable and prone to systematic biases in the absence of true historical information. Hansen et al. (2008) presented another approach to deal with this problem based on assuming a model of evolution for the predictor variables and then using only tip-species observations as data. Their model relied on Brownian motion for the predictors and is thus only applicable to continuous predictor variables, although is possible to combine this with discrete reconstructed niches in ANCOVA types of models for the primary optimum (Escudero et al. 2012; Voje and Hansen 2013). Extensions to randomly

evolving discrete predictors (e.g., discrete niches) based on simple Markov chain models are possible, but have not yet been developed. Bartoszek et al. (2012) extended the method to multivariate response variables that can both influence each other's primary optima and be stochastically correlated. A special case of the Bartoszek et al. (2012) model extends Hansen et al. (2008) by allowing predictor variables to follow an Ornstein–Uhlenbeck model. Beaulieu et al. (2012) extended the method in a different direction by allowing not only the optima but also the α-parameter and σ-parameter to differ among reconstructed niches. This allows testing of niche- and clade-dependent rates of adaptation and evolution in a quite general and flexible manner (Chap. 15). See also Lajeunesse (2009) for implementation of the method into a meta-analytic framework.

It is important to understand that little is gained by just using the Ornstein–Uhlenbeck process to model the variances and covariances in the residual **V** matrix as done in many studies and implemented in some software. This simply transforms the form of the relationship between covariance and phylogeny and does not capture the effects of species tracking optima throughout the phylogeny. It does not adequately address any of the above-mentioned problems, which derive from the artificial separation of adaptation and residual deviation.

14.7   Using the Method to Study Adaptation of Deer Antler Size

For illustration, I present some analyses of a data set on deer antler size from Plard et al. (2011) and also analyzed by Bartoszek et al. (2012). The basic data are measures of antler length in 32 species from the deer family (Cervidae), and to test the hypothesis that antler size is influenced by sexual selection, we use a classification from Clutton-Brock et al. (1980) of the mating system into large, medium, and small breeding groups. The idea is that there will be stronger sexual selection in the large breeding groups. The small breeding group is close to monogynous systems. In Fig. 14.6a, we see a parsimony mapping of the breeding-group systems on a phylogeny. I used the program Slouch (Hansen et al. 2008) to run an analysis with log antler length as response variable and breeding-group-size niche as a predictor variable with three states. The estimated primary optima for antler size corrected for body mass are shown in Fig. 14.6b. They indicate a strong effect in the predicted direction, and the results are consistent with sexual selection being important. The phylogenetically corrected R^2 is 41 % after body size is corrected for and 87 % for the whole model.

These estimates are conditional on the best estimates of the α-parameter and σ-parameter of the Ornstein–Uhlenbeck model. Figure 14.6c shows the likelihood surface for these parameters that to aid interpretation have been transformed to two new parameters: $t_{1/2} = \ln2/\alpha$ and $v_y = \sigma^2/2\alpha$. The phylogenetic half-life, $t_{1/2}$, is the average time it takes to evolve half the distance from the ancestral state towards an
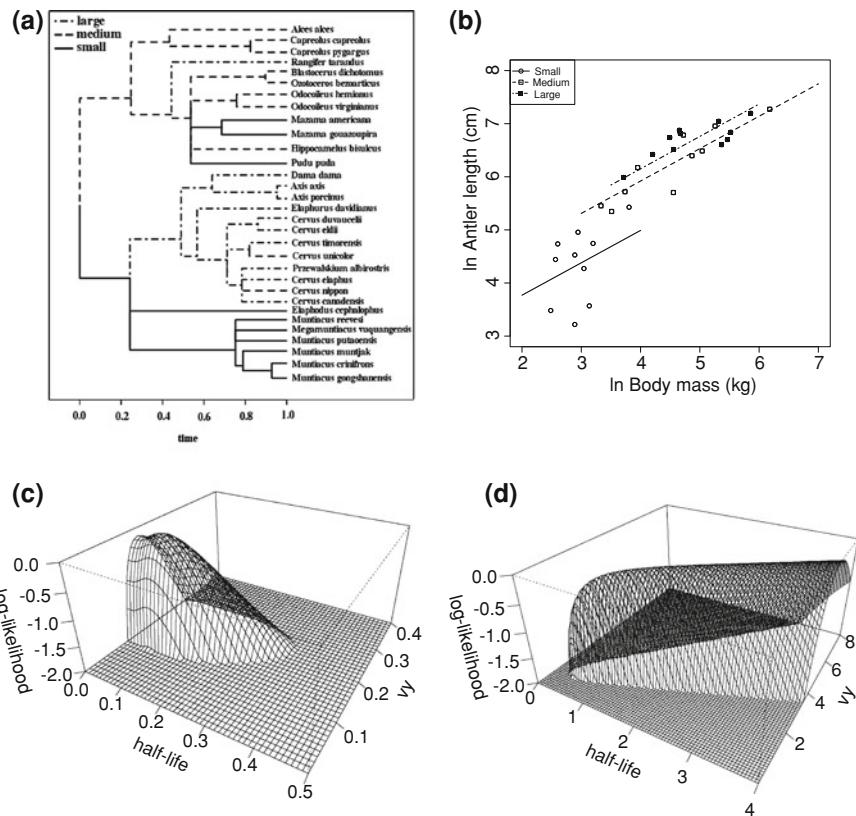
Fig. 14.6 Analyzing adaptation of antler size to breeding-group size using Slouch. **a** Mapping of *small*, *medium*, and *large* breeding groups as "niches" on the phylogeny (reprinted from Bartoszek et al. 2012 with permission from Elsevier). **b** Estimated primary optima for log antler length on the three breeding group sizes given as regressions on log body mass. The slope on log body mass is 0.61 ± 0.18 ($\pm$standard error), and the intercepts for the *small*, *medium*, and *large* breeding groups are, respectively, 2.55 ± 0.56, 3.48 ± 0.83, and 3.71 ± 0.90. The model explains $R^2 = 87$ % of the variance. **c** Log-likelihood surface for phylogenetic half-life $t_{1/2}$ and stationary variance, v_y for this model. The best estimates are $t_{1/2} = 0.07$th and $v_y = 0.16$. **d** Log-likelihood surface for the same parameters from a model with only an intercept. The best estimates here are $t_{1/2} = 2.4$th and $v_y = 3.6$, but the surface has a ridge that extends outward to infinity indicating that the phylogenetic effect is indistinguishable from the pattern expected from Brownian motion. Note different scales from **c**. Flat areas in the likelihood surfaces correspond to parameter values for which the log-likelihood is more than two units worse than the best estimate. Data for analysis originally from Plard et al. (2011)

optimum in a new niche as illustrated in Fig. 14.5b. The stationary variance, v_y, is a measure of how much the trait tends to deviate from the primary optimum when evolution has come to a stochastic equilibrium. The likelihood function peaks at a half-life of $t_{1/2} = 7$ % of time from tip to root of the phylogenetic tree. Hughes et al. (2006) estimated that the most recent common ancestor of the Cervidae lived

between 16 and 23 million years ago. Taking this into account, the half-life is between $t_{1/2} = 1.1$ and $t_{1/2} = 1.6$ million years. While this corresponds to a considerable lag in adaptation from a microevolutionary perspective, it is practically undistinguishable from instant adaptation on the timescales of this phylogeny. Instantaneous adaptation ($t_{1/2} = 0$) has only marginally lower likelihood. We can exclude long half-lives, however. The log-likelihood has dropped about two units when the half-life has reached 30 % of the distance from root to tip, which may be described as a moderate phylogenetic effect.

Now consider Fig. 14.6d, which shows the log-likelihood surface for $t_{1/2}$ and v_y for a model with only an intercept (i.e., no predictor). In this case, the phylogenetic half-life measures the phylogenetic signal in the trait (log antler size) as opposed to in the model residuals. The best estimate of the half-life in this case is 2.4 times the time back to the most recent common ancestor of the Cervidae, or between 38 and 55 million years. Half-lives that are much longer than the time back to the common ancestor are essentially indistinguishable from the pattern of a Brownian motion, which has a half-life of infinity. Consequently, the likelihood surface in Fig. 14.6d has a long ridge that extends out to infinity with only a microscopic drop in likelihood. In contrast, zero half-life is here soundly rejected (about 8 log-likelihood units worse). Hence, if we look at antler size in isolation, we see a pattern that resembles Brownian motion, and any conventional check for phylogenetic signal would be consistent with the use of a standard phylogenetic method as the independent contrasts. Looking at the phylogenetic signal in model residuals in Fig. 14.6c, however, shows that this would be a serious mistake. The weak phylogenetic signal in the residuals means that a non-phylogenetic analysis would be close to optimal. The reason why this happens is that antler size tracks predictor variables with strong phylogenetic effects (as seen in Fig. 14.6a) and thus inherits this phylogenetic signal even if there is little phylogenetic signal in the residual deviations. This is precisely the situation that was illustrated in Fig. 14.3a.

Although consistent with a strong effect of sexual selection on antler size, the fitted model is crude. The characterization of the breeding group niches is qualitative, the reconstruction of niches on the phylogeny is almost certainly seriously inaccurate, the phylogeny itself is inaccurate (cmp. Figure 14.2 with Fig. 14.6a), and the antler- and body-size measurements have unknown measurement error. It is indeed remarkable that the model explains so much of the variance. Bartoszek et al. (2012) give a more detailed analysis of the same data. There, antler size and sexual size dimorphisms are analyzed together as coevolving variables based on a multivariate Ornstein–Uhlenbeck process around niche-dependent optima. This analysis shows that antler size and sexual size dimorphisms are not coevolving in the sense of influencing each others' primary optima, but they show correlated evolution in their stochastic deviances that must be due to correlations in some secondary variables, and their primary optima depend on breeding group size in the predicted manner.

14.8 Interpreting the Parameters
in the Ornstein–Uhlenbeck Model

Lande (1976) showed that the Ornstein–Uhlenbeck process could be derived from a simple model of quadratic stabilizing selection and genetic drift, and this is sometimes used to justify and interpret the parameters in the comparative method. Under this interpretation, the α-parameter becomes equal to the product of additive genetic variance and the curvature of the stabilizing selection function and the σ^2 parameter becomes equal to the additive variance divided by the effective population size. This is on a generational timescale, however, and on the million-year timescales that usually separate distinct species, the rates of evolution that can be predicted from estimates of evolvability and strengths of selection are so high as to be essentially instantaneous (Hansen 2012).

The conclusion from this is that evolution on a fixed fitness surface is usually so rapid that we do not expect any ancestral effects or any similarity of related species beyond what is due to similar positions of their adaptive optima. We know, however, that related species are similar, and this requires an explanation. A clue to explanation may be to ask why a primary optimum may not be reached instantaneously.

A possibility suggested by Hansen (1997) was based on Simpson's (1944) idea that primary adaptations may be slow due to the need to establish a number of secondary adaptations before it can reach its fullest expression. This idea is perhaps most clearly expressed by Kemp (2006, 2007), who referred to it as the correlated-progression hypothesis. Here, we imagine the focal biological trait as embedded in a complex network of coadapted interactions with other traits. Any large change in a focal trait is thus unlikely to be beneficial due to internal selective constraints even if there is external selection to change it. A small change may, however, be beneficial on balance, and this will set up selection on other trait to adjust to this small change, which will again make possible a further change in the focal trait, and so on. Thus, external directional selection on a trait (or suite of traits) may result in a slow correlated progression. On this view of evolution, a small α-parameter may result from strong internal selective constraints. This point is discussed in more detail in Labra et al. (2009) and Hansen (2012).

The definition of the primary optimum as an average optimum over repeated reruns of evolution in a given niche assumes that evolution in the niche tends to evolve around this optimum. It does not capture cases with two or more distinct centers of attraction, as with distinct alternative strategies or disruptive selection where more extreme alternatives are favored. We may, however, extend the concept to cover niche-dependent trends of evolution. A primary trend may be defined as the average directional rate of evolution over repeated reruns of evolution in the given niche. In the Ornstein–Uhlenbeck model, this manifests itself in cases where a long phylogenetic half-life is combined with primary optima that are far outside the range of among-species variation. In this case, the average directional change per time is equal to $\alpha\theta$, and the trends can thus be estimated by

multiplying the estimated primary optima with the estimated α-parameter. In these situations, the model behaves like a Brownian motion with niche-dependent trends (see Hansen 1997).

14.9 Software and Applications

These adaptation methods have now been used in many studies and are at least partially implemented in several software packages. The popularity of the methods is to a large degree due to Butler and King's (2004) R package Ouch, which can estimate optima in fixed niches mapped onto a phylogeny and assess the fit of different niche arrangements. Later, several other packages with complementary functionality have been developed. Slouch, originally developed by Hansen et al. (2008) to fit the model with randomly evolving predictor variables, has also been extended to handle fixed niches mapped on the phylogeny combined with random effects. It differs from Ouch by using different parameterizations of the model ($t_{1/2}$ and v_y in place of α and σ^2) and by using a grid search in place of a numerical optimization algorithm. The latter is an advantage in that it makes the uncertainty in parameter estimation more apparent and guards against convergence on suboptimal peaks in the likelihood, but a disadvantage with large numbers of species, as it can be slow and cumbersome to use. Slouch can also be used to correct for known observation variance in both response and predictor variables. Beaulieu et al's (2012) Ouwie allows not only the primary optima but also the α-parameter and σ-parameter to depend on niches, and is hence useful for estimating environment-dependent rates of evolution and adaptation (in this respect, it extends Brownie, O'Meara et al. 2006). MvSlouch (Bartoszek et al. 2012) allows several traits to evolve around a multivariate primary optimum determined by fixed niches or other randomly evolving variables. A restricted multivariate model can also be fitted by Ouch. The multivariate models are very parameter rich, however, and their application is technically difficult and requires much data and well-specified hypotheses to constrain the number of parameters. The program Surface (Ingram and Mahler 2013) includes routines for identifying and comparing models of convergent evolution based on alternative niche reconstructions on the phylogeny (Chap. 18). Versions of the model can also be fit by the programs Ape (Paradis et al. 2004), Brownie (O'Meara et al. 2006), Compare (Martins 2004), and Geiger (Harmon et al. 2008).

I am aware of more than 70 published applications of these methods, which will be reviewed elsewhere (Hansen, Pienaar, Voje, Bartoszek in preparation). The main impression is that most of these studies are able to find evidence of adaptation in that they can convincingly reject alternative non-adaptive models and provide parameter estimates in accordance with a priori expectations from the adaptive hypothesis. Importantly, adaptation is also sometimes rejected. For example, Labra et al. (2010) found no signs that the size of the parietal (third) eye was adapted to any climatic or thermophysiological variables in a lizard genus (*Liolaemus*), thereby providing evidence against long-standing hypotheses of

thermobiological functionality for this organ. It is commonly found that adaptation is fast enough to produce none to mild residual phylogenetic correlations, but this is not always the case, and it remains necessary to consider phylogeny. The standard phylogenetic comparative model would, however, only have been appropriate in a small minority of cases and would then also need to be reinterpreted in terms of estimating trends. A recurring problem with many of these studies is the failure to report parameter estimates, or reporting them without units. This limits the possibility for interpreting and generalizing from this body of work.

14.10 Testing for Adaptation in Qualitative (Categorical) Traits

The study of adaptation has always been one of the major uses for comparative methods. Most usage is simply looking for predicted associations between traits and environments without controlling for phylogeny, or based on phylogenetic corrections that are not justified with appropriate process models. Either of these approaches may be informative in a qualitative sense, but falls short of delivering logically consistent quantitative estimates of precisely defined parameters. For continuous quantitative characters, the explicit adaptation models are now sufficiently well developed to fill the need.

The situation is less clear for qualitative traits. Although there are stochastic-process models appropriate for categorical traits, there has been no discussion of how to parameterize adaptation along the lines I discussed above. Simple Markov-chain models can be used to estimated the intensities (probabilities per time) of change between states of a variable as a function of the state of another variable (e.g., Pagel 1994), and this allows tests of adaptation in similar ways as above (e.g., Hansen and Orzack 2005). In particular, maintenance of a trait A in the presence of environment B can be used as evidence of adaptation of trait A to environment B. Armbruster (2002) used this method to test whether bract color (green/white vs. pink/purple) in *Dalechampia* blossoms depended on type of bee pollinator (euglossine vs. megachilid). He found no support for this in a study of 37 species. He did, however, find support for an influence of vegetative colors on the evolution of bract colors, concluding that bract colors may be changing due to indirect selection stemming from a pleiotropic relation to stem pigments.

In analogy with the adaptation model for continuous traits, one could define adaptation as a systematic niche-dependent effect on the rates of transition between states, or equivalently, on the equilibrium probabilities of the states. This could be done directly in a generalized linear model setup or indirectly by allowing the categorical trait to be a threshold function of an underlying continuous variable that itself follows the adaptation model, but such models remain to be developed. Felsenstein (2012) has shown how a threshold model on traits evolving as Brownian motion can be fit to discrete comparative data, and it should be possible to extend this to Ornstein–Uhlenbeck models (see also Chap. 16).

14.11 Ancestral-State Reconstruction and Inferred-Changes Methods

Perhaps the most common comparative test of adaptation is to reconstruct character and environmental states on a phylogeny and then use these as data in various statistical analyses. In the cladistic tradition, this is done by inferring trait changes with parsimony methods and then using these inferred changes as data in subsequent analyses (Maddison 1994). Ancestral states and changes can also be inferred by assuming specific evolutionary process models such as Brownian motion or Markov chains (Martins and Hansen 1997; Schluter et al. 1997) and then either the reconstructed states at nodal values or the inferred changes along branches can be used as data in subsequent analyses.

Unfortunately, inferred changes on a phylogeny are poorly suited for statistical analysis (Frumhoff and Reeve 1994). In the cladistic tradition, inferred changes are treated as independent evolutionary events, and it is assumed that using them as data solves the phylogenetic correlation problem that arises when using species data. Ridley (1983, p. 18) makes this argument explicit when he writes "To recognize independent evolution is to distinguish primitive from derived character states. Derived characters are independently evolved characters." From this, he proposed a research strategy for testing adaptation based on using inferred changes from primitive to derived character states as independent data in statistical analyzes. In my opinion, his argument is mistaken in two interesting ways.

The first and most obvious problem is that the inferred changes to be used as data are not the actual changes that have happened in the history of life, but estimates thereof. Even if the true changes could be regarded as independent evidence, the estimates cannot. Inferred changes based on parsimony (or any other method) are not independent of each other. Whether a change is inferred to have happened on a particular branch in the phylogeny depends on what changes are inferred on other branches. Using inferred changes as data in non-phylogenetic statistical analysis will seriously violate the assumptions of independent sampling and is likely to give misleading information about evolution.

If the ancestral states are inferred from a statistical model such as a Markov chain or a Brownian motion, then it is possible to derive the joint statistical distribution of the inferred changes including their variances and covariances (Martins and Hansen 1997). This makes it possible to use them correctly in statistical analyses that take account of covariance and heteroscedacity. Doing this, however, reveals a fundamental circularity of the inferred-changes approach. First, a model is assumed and parameters may be estimated to infer the changes, and then, the inferred changes are used to make inferences about evolution. It is clear that those inferences are constrained to reflect the assumptions of the model used to infer the changes. For example, a set of ancestral states reconstructed from extant species data by assuming a Brownian-motion process will have the same average as the extant species. Hence, one will observe that there is no temporal trend in the (inferred) data, but this is just recovering a property of the assumed

model and does not constitute evidence against evolutionary trends. More subtle errors of this sort are hard to avoid. In any case, the likelihood principle of statistics tells us that given an evolutionary process model, then all the information about its parameters is contained in the probability of the observed data given the model (i.e., in the likelihood function). From this, it follows that reconstruction of ancestral states can never provide any information that cannot be obtained by a standard likelihood analysis of the observed tip species. The best that can be hoped from an inferred-changes method is a roundabout way to construct a likelihood function, and I am not aware of a single application that has achieved even that.

In addition to the statistical problem of using inferred changes as if they were true changes, there is also an interesting biological problem with Ridley's reasoning. It concerns his assumption that the true changes from primitive to derived character states can be treated as independent observations of an evolutionary process. The problem with this can be seen from our model (14.1) of how a trait is determined by a number of observed and unobserved factors. The model was: $Y = b_0 + b_1 X_1 + r(X_1, \ldots, X_m)$. If a change on a branch happens because the focal variable, X_1, changes with an amount ΔX_1, then we may predict a change $\Delta Y = b_1 \Delta X_1$, which would be independent of changes on other branches provided, and this is the problem, provided that there are no correlations between branches in the changes of secondary variables that make the residuals. If, due to some inherited common biology, X_2 has a tendency to increase in one part of the phylogeny, then the residuals from a regression of ΔY on ΔX_1 will tend to be similar in this part of the phylogeny. Put in other words, evolutionary changes in different parts of a phylogeny are expected to be correlated for exactly the same reason that species trait values in different parts of the phylogeny are expected to be correlated. The underlying reason is that both states and changes are dependent on shared third variables inherited from a common ancestor. This is also the reason why pairwise contrasts on a phylogeny cannot be taken to be independent except under specified precise assumptions about the underlying evolutionary process, such as in the derivation of independent contrasts from a multivariate Brownian motion. For example, it can be shown that independent contrasts are not independent if the species data were generated by an Ornstein–Uhlenbeck process.

In molecular evolution, the reconstruction of ancestral genes and proteins is becoming increasingly popular. This approach has an advantage over the reconstruction of higher level physiological, morphological, behavioral, and life-history traits in that molecular reconstructions are based on models of evolution that are better understood and delineated. Still, the fundamental problems are the same, and reconstructed genes and proteins should not be used to test laws of adaptation. Also here, the principled approach would be to estimate parameters of defined models based on extant species data without the vivid but statistically unnecessary pass through a reconstructed ancestor.

14.12 A Measurement-Theoretical Perspective
on Comparative Methods

Measurement is the assignment of numbers to attributes of reality. Measurements are valid when relations among the numbers reflect relevant relations among the attributes of reality. Only then can inferences made from the numbers lead to meaningful inferences about nature. Measurement theory is the study of the mapping from reality to numbers. Houle et al. (2011) pointed out that a measurement-theoretical perspective has been lacking from biology and argued that a large number of common problems, ranging from relatively obvious errors such as ignoring units or using p values as measures of effect to more subtle problems involving violations of scale type or making statistical manipulations that violate the theoretical context of the study, can be diagnosed as violation of measurement-theoretical principles.

Comparative studies are extremely vulnerable to such problems because there is a need to simultaneously handle difficult statistical problems, error prone data of heterogeneous origin, and quantification of poorly understood theories of evolution on long timescales. The drive to solve statistical problems can easily come into conflict with studying meaningful biological relations, and without a firm grasp of the meaning of measurement, it is easy to loose the biological baby with the statistical bathwater.

An undercurrent of this essay has been that the statistical laundering that took place with the development of phylogenetic comparative methods has also weakened the connections to biology, and particularly so in the treatment of adaptation. I have argued that the assumption of evolution as Brownian motion has fundamental incompatibilities with adaptation in the sense of evolution toward niche-dependent optima. The Brownian-motion assumption arose from the need to model interspecies covariances as shared branch lengths so as to be able to use GLS techniques without having to estimate extra nuisance parameters in the variance matrix. This implicitly assumes that the effects of adaptation could be separated from the residual covariances and that the estimated parameters in regression and ANOVAs could be interpreted in the same way as they would in non-phylogenetic analyses. As shown above, the meaning of these parameters change. For example, the regression slope becomes influenced both by the optimal relation and by the expected evolutionary lag (Fig. 14.4).

Evolutionary regression studies are often conducted in a descriptive statistical manner without quantitative links to the theory that motivated the study. A qualitative prediction of, say, a positive relationship between two variables may motivate the study, but the context of this prediction is often not used to inform or constrain the statistical procedure. When statistical manipulations are decoupled from theoretical context, this will often lead to measurement-theoretical problems. Using interspecies trait covariances in place of residual covariances needs justification, and even if such justification can be derived in some cases, as when all variables follow a joint multivariate Brownian motion, it is typically incorrect, as

discussed above. I suspect this mistake was allowed to proliferate to such a degree because there was no demand to formally justify statistical methods and procedures from biological assumptions. I will now briefly discuss two more examples where statistical arguments void of biological justification have lead to serious errors in comparative studies.

One case concerns the estimation of allometric relationships, which is important both as a goal in itself by testing hypotheses about scaling relationships and functional adaptations, and due to its widespread use in controlling for body size. While essentially all theory about allometric scaling refers to a power relation between variables (i.e., $Y = aX^b$, where Y is a trait and X is usually a measure of body size, as in Fig. 14.1), there appeared in the 1970s a verbal broadening of the term allometry to describe all kinds of nonlinear trait relationships. This lead to a curious situation where predictions derived from or about allometry in the narrow sense of a power relationship were tested with statistical methods incompatible with the power relationship. For example, the allometric exponent b is traditionally estimated from a log–log regression ($\log[Y] = \mathrm{Log}[a] + b\mathrm{Log}[X]$), but in the context of broad-sense allometry, the exponent has no meaning and the theoretical necessity of the log-transformation disappears, so that researchers started to do "allometric" linear regression on the arithmetic scale (see Houle et al. 2011; Voje et al. 2014 for discussion and documentation). This was sometimes accompanied with arguments that the log-transformation was not necessary for statistical reasons (e.g., it was not necessary to stabilize the variance or to achieve normality). When theoretical context was thus forgotten, a theoretically necessary manipulation became confused with a statistical manipulation and was then often dropped for statistical reasons. One result of this was a substantial body of work and an emerging consensus that static and ontogenetic "allometries" are evolvable and flexible traits and not important constraints on evolution. Voje et al. (2014; Voje and Hansen 2013) have shown, however, that none of this work apply to allometry in the traditional narrow sense and that there is in fact considerable evidence that static allometries are highly constrained and may therefore act as evolutionary constraints on trait adaptation even on macroevolutionary timescales as hypothesized by Huxley (1932), Rensch (1959), Gould (2002), and others.

Another common violation of measurement-theoretical principles in cross-species regression studies involves the use of nonparametric curve-fitting techniques such as reduced major-axis regression. Such techniques are based on geometrical arguments on how to draw a line through points and not justified in terms of estimating parameters in a defined model. In particular, the common practice of using the reduced major-axis slope to estimate allometric exponents has obscured our empirical knowledge of allometric scaling relationships (Voje et al. 2014). This slope is computed as the ratio of the standard deviations of the variables involved and does not even include their covariance. Except when points deviate little from a straight line, it typically gives a seriously misleading estimate of a true underlying regression slope (Kelly and Price 2004; Hansen and Bartoszek 2012). Using the reduced major-axis slope to estimate an allometric exponent or a causal effect of a predictor variable on a trait are measurement-theoretical mistakes

because there are no connections between the statistical procedure and the entities to be estimated.

Reduced major-axis and other nonparametric regression techniques are sometimes said to improve on standard regression when there is "error" in both response and predictor variables. This is only true under highly specific and unlikely circumstances and seems to derive from an implicit assumption that all residual deviations are due to measurement error in the variables and not to biological deviations from the model (Hansen and Bartoszek 2012). In reality, most residual deviations from evolutionary regression models are due to biological deviations from the linear relationship as described in Eq. (14.1). Indeed, if this was not the case, there would be no reason to worry about phylogenetic correlations! In such cases, the "corrections" of the major axis and more general structural equation models are completely off the mark. In particular, reduced major-axis regression should never be used in comparative analyses, and every result drawn from this method should be reconsidered.

Yet, this is not to deny that measurement error is a major concern in comparative analyses. In a typical comparative analysis, data have to be gathered from many individuals within each of many species in a consistent manner. Often sample sizes for individual species are small and vary with orders of magnitude between species, so that some are much more reliable than others (Garamszegi and Møller 2010; 2011). Many comparative studies are also based on compiled data sets of dubious quality. For example, the data on antler-body-size regression shown in Fig. 14.1 and used in the classical analyses by Gould (1974) and Clutton-Brock et al (1980) are reported without any indication of uncertainty and appear to be of poor quality. As just one example, I traced the estimate of the body size of the fallow deer given as a 91.0-cm shoulder height to the following statement in Ward's (1903) *Record of big game*: "Height at shoulder about 3 feet" (Ward 1903, p. 64). Somehow this qualitative statement transmuted into a fixed quantitative measurement with three significant digits on the cm scale ready to be used in the comparative studies. This may seem outrageous, but I fear it is quite typical. Smith and Jungers (1997) documented even worse problems in a standard data set of primate body size used in many comparative studies.

The most important thing to do about this is to pay more attention to data quality and also not to accept measurements without units and indications of uncertainty (as I admittedly did in my example). If estimates of measurement error variation are available, for example, as standard errors of species means, methods have been developed to include these into phylogenetic comparative studies (e.g., Lynch 1991; Martins and Hansen 1997; Ives et al. 2007; Felsenstein 2008; Lajeunesse 2009; Hadfield and Nakagawa 2010; Hansen and Bartoszek 2012; Revell and Reynolds 2012; reviewed in Chap. 7). Also the bias in regression slopes induced by measurement error in predictor variables that sometimes motivates the use of nonparametric regression can easily be corrected with simple extensions of the standard phylogenetic linear model (Hansen and Bartoszek 2012).

14.13 Conclusion

The emergence of phylogenetic comparative methods may appear a textbook case of scientific progress. The growing recognition of statistical problems and inconsistencies led to increasing dissatisfaction with comparative biology, but this was resolved by the development of new methods of measurement and the incorporation of new types of data in the form of molecular phylogenies. Phylogenetic comparative methods then became the core of a new research paradigm that was more quantitative, statistical, and rigorous than what went before. Here, I have argued that this paradigm shift was not without its problems. In relation to the study of adaptation, the solutions to the statistical problems generated conceptual and interpretational inconsistencies that need to be resolved. I believe these inconsistencies at least now have been identified and that solutions are emerging at an accelerating pace. There is still more work to be done before we are able to deal with all the different types of data and theories that we would like to include in our analyses and tests of adaptation with comparative methods, but I hope this chapter can help clarify the principles for how this can be done.

References

Armbruster WS (2002) Can indirect selection and genetic context contribute to trait diversification? A transition-probability study of bolssom-colour evolution in two genera. J Evol Biol 15:468–486

Bartoszek K, Pienaar J, Mostad P, Andersson S, Hansen TF (2012) A comparative method for studying multivariate adaptation. J Theor Biol 314:204–215

Baum DA, Larson A (1991) Adaptation reviewed: a phylogenetic methodology for studying character macroevolution. Syst Zool 40:1–18

Beaulieu JM, Jhwueng D-C, Boettiger C, O'Meara BC (2012) Modeling stabilizing selection: expanding the Ornstein-Uhlenbeck model of adaptive evolution. Evolution 66:2369–2383

Brooks DR, McLennan DH (1991) Phylogeny, ecology and behavior: a research program in comparative biology. University of Chicago Press, Chicago

Butler MA, King AA (2004) Phylogenetic comparative analysis: a modeling approach for adaptive evolution. Am Nat 164:683–695

Cheverud JM, Dow MM, Leutenegger W (1985) The quantitative assessment of phylogenetic constraints in comparative analyses: sexual dimorphism in body weights among primates. Evolution 39:1335–1351

Clutton-Brock TH, Albon SD, Harvey PH (1980) Antlers, body size and breeding group size in the Cervida. Nature 285:565–567

Coddington JA (1988) Cladistic tests of adaptational hypotheses. Cladistics 4:3–22

Cooper N, Jetz W, Freckleton RP (2010) Phylogenetic comparative approaches for studying niche conservatism. J Evol Biol 23:2529–2539

Diniz-Filho JAF, Ramos de Sant'ana CE, Bini LM (1998) An eigenvector method for estimating phylogenetic inertia. Evolution 52:1247–1262

Diniz-Filho JAF, Rangel TF, Santos T, Bini LM (2012) Exploring patterns of interspecific variation in quantitative traits using sequential phylogenetic eigenvector regressions. Evolution 66:1079–1090

Escudero M, Hipp A, Hansen TF, Voje KL, Luceño M (2012) Selection and inertia in the evolution of holocentric chromosomes in sedges (*Carex,* Cyperaceae). New Phytol 195:237–247

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

Felsenstein J (2008) Comparative methods with sampling error and within-species variation: contrasts revisited and revised. Am Nat 171:713–725

Felsenstein J (2012) A comparative method for both discrete and continuous characters using the threshold model. Am Nat 179:145–176

Freckleton RP, Cooper N, Jetz W (2011) Comparative methods as a statistical fix: the dangers of ignoring an evolutionary model. Am Nat 178:E10–E17

Frumhoff PC, Reeve HK (1994) Using phylogenies to test hypotheses of adaptation: a critique of some current proposals. Evolution 48:172–180

Garland T Jr, Bennett AF, Rezende EL (2005) Phylogenetic approaches in comparative physiology. J Exper Biol 208:3015–3035

Garamszegi LZ, Møller AP (2010) Effects of sample size and intraspecific variation in phylogenetic comparative studies: a meta-analytic review. Biol Rev 85:797–805

Garamszegi LZ, Møller AP (2011) Nonrandom variation in within-species sample size and missing data in Phylogenetic comparative studies. Syst Biol 60:876–880

Geist V (1998) Deer of the world: their evolution, behavior, and ecology. Swan Hill Press, Shrewsbury

Gould SJ (1973) Positive allometry of antlers in the "Irish elk", *Megaloceros giganteus*. Nature 244:375–376

Gould SJ (1974) The evolutionary significance of "bizarre" structures: antler size and skull structure in the "Irish Elk," *Megaloceros giganteus*. Evolution 28:191–220

Gould SJ (1977) Ontogeny and Phylogeny. Belknap, Cambridge

Gould SJ (1998) A lesson from the old masters. In: Gould SJ (ed) Leonardo's mountain of clams and the diet of worms. Harmon books, pp 179-196

Gould SJ (2002) The structure of evolutionary theory. Harvard University Press, Cambridge

Gould SJ, Vrba ES (1982) Exaptation—a missing term in the science of form. Paleobiology 8:4–15

Hadfield JD, Nakagawa S (2010) General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. J Evol Biol 23:494–508

Hansen TF (1997) Stabilizing selection and the comparative analysis of adaptation. Evolution 51:1341–1351

Hansen TF (2012) Adaptive landscapes and macroevolutionary dynamics. In: Svensson EI, Calsbeek R (eds) The adaptive landscape in evolutionary biology. Oxford University Press, Oxford, pp 205–226

Hansen TF, Bartoszek K (2012) Interpreting the evolutionary regression: the interplay between observational and biological errors in phylogenetic comparative studies. Syst Biol 61:413–425

Hansen TF, Martins EP (1996) Translating between microevolutionary process and macroevolutionary patterns: the correlation structure of interspecific data. Evolution 50:1404–1417

Hansen TF, Orzack SH (2005) Assessing current adaptation and phylogenetic inertia as explanations of trait evolution: the need for controlled comparisons. Evolution 59:2063–2072

Hansen TF, Pienaar J, Orzack SH (2008) A comparative method for studying adaptation to a randomly evolving environment. Evolution 62:1965–1977

Harmon LJ, Weir JT, Brock CD, Glor RE, Challenger W (2008) Geiger: investigating evolutionary radiations. Bioinformatics 24:129–131

Houle D, Pélabon C, Wagner GP, Hansen TF (2011) Measurement and meaning in biology. Quart Rev Biol 86:3–34

Hughes S, Hayden TJ, Douady CJ, Tougard C, Germonpre M, Stuart A, Lbova L, Carden RF, Hanni C, Say L (2006) Molecular phylogeny of the extinct giant deer, *Megaloceros giganteus*. Mol Phylogenet Evol 40:285–291

Huxley JS (1932) Problems of relative growth. Lincoln Mac Veagh-The Dial Press, New York

Ingram T, Mahler DL (2013) SURFACE: detecting convergent evolution from comparative data by fitting Ornstein-Uhlenbeck models with stepwise Akaike information criterion. Methods Ecol Evol 4:416–425

Ives AR, Garland T Jr (2010) Phylogenetic logistic regression for binary dependent variables. Syst Biol 59:9–26

Ives AR, Midford PE, Garland T Jr (2007) Within-species variation and measurement error in phylogenetic comparative methods. Syst Biol 56:252–270

Kelly C, Price TD (2004) Comparative methods based on species mean values. Math Biosci 187:135–154

Kemp TS (2006) The origin of mammalian endothermy: a paradigm for the evolution of complex biological structure. Zool J Linn Soc 147:473–488

Kemp TS (2007) The origin of higher taxa: macroevolutionary processes, and the case of the mammals. Acta Zoologica 88:3–22

Labra A, Pienaar J, Hansen TF (2009) Evolution of thermal physiology in *Lioleamus* lizards: adaptation, phylogenetic inertia and niche tracking. Am Nat 174:204–220

Labra A, Voje KL, Seligmann H, Hansen TF (2010) Evolution of the third eye: a phylogenetic comparative study of parietal-eye size as an ecophysiological adaptation in *Liolaemus* lizards. Biological J Linn Soc 101:870–883

Lajeunesse MJ (2009) Meta-analysis and the comparative phylogenetic method. Am Nat 174:369–381

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

Larson A, Losos JB (1996) Phylogenetic systematics of adaptation. In: Rose MR, Lauder GW (eds) Adaptation. Academic press, San Diego, pp 187–220

Lister AM et al (2005) The phylogenetic position of the 'giant deer' *Megaloceros giganteus*. Nature 438:850–853

Lynch M (1991) Methods for the analysis of comparative data in evolutionary biology. Evolution 45:1065–1080

Maddison DR (1994) Phylogenetic methods for inferring the evolutionary history and processes of change in discretely valued characters. Ann Rev Entomol 39:267–292

Martins EP (2000) Adaptation and the comparative method. Trends Ecol Evol 15:296–299

Martins EP (2004) Compare, version 4.6b. Computer programs for the statistical analysis of comparative data. Distributed by the author at http://compare.bio.indiana.edu/. Technical report, Department of Biology, Indiana University, Bloomington, IN

Martins EP, Hansen TF (1997) Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. Am Nat 149:646–667

Mitchell WA, Valone TJ (1990) The optimization research program: studying adaptations by their function. Quart Rev Biol 65:43–52

Nunn CL (2011) The comparative approach in evolutionary anthropology and biology. The University of Chicago Press, Chicago

O'Meara BC (2012) Evolutionary inferences from phylogenies: a review of methods. Annu Rev Ecol Evol Syst 43:267–285

O'Meara BC, Ane C, Sanderson MJ, Wainwright PC (2006) Testing for different rates of continuous trait evolution using likelihood. Evolution 60:922–933

Pagel MD (1994) Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proc R Soc B 255:37–45

Paradis E, Claude J, Strimmer K (2004) Ape: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290

Pennell MW, Harmon LJ (2013) An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology. Ann New York Acad Sci 1289:90–105

Plard F, Bonenfant C, Gaillard J-M (2011) Revisiting the allometry of antlers among deer species: male–male sexual competition as a driver. Oikos 120:601–606

Price T (1997) Correlated evolution and independent contrasts. Philos Trans R Soc B 355:1599–1606

Reeve HK, Sherman PW (1993) Adaptation and the goals of evolutionary research. Quart Rev Biol 68:1–32

Reeve HK, Sherman PW (2001) Optimality and phylogeny: a critique of current thought. In: Orzack SH, Sober E (eds) Adaptationism and optimality. Cambridge University Press, Cambridge, pp 64–113

Rensch B (1959) Evolution above the species level. Wiley, New York

Revell LJ (2010) Phylogenetic signal and linear regression on species data. Methods Ecol Evol 1:319–329

Revell LJ, Reynolds G (2012) A new bayesian method for fitting evolutionary models to comparative data with intraspecific variation. Evolution 66:2697–2707

Ridley M (1983) The explanation of organic diversity: the comparative method and adaptations for mating. Clarendon Press, Oxford

Schluter D, Price T, Mooers AØ, Ludwig D (1997) Likelihood of ancestor states in adaptive radiation. Evolution 51:1699–1711

Simpson GG (1944) Tempo and mode in evolution. Columbia University Press, New York

Smith RJ, Jungers WL (1997) Body mass in comparative primatology. J Human Evol 32:523–559

Sober E (1984) The nature of selection: evolutionary theory in philosophical focus. Bradford books, Cambridge

Sober E (2008) Evidence and evolution: the logic behind the science. Cambridge University press, Cambridge

Stone GN, Nee S, Felsenstein J (2011) Controlling for non-independence in comparative analysis of patterns across populations within species. Phil Trans R Soc B 366:1410–1424

Voje KL, Hansen TF (2013) Evolution of static allometries: slow rate of adaptive change in allometric slopes of eye span in stalk-eyed flies. Evolution 67:453–467

Voje KL, Hansen TF, Egset CK, Bolstad GH, Pélabon C (2014) Allometric constraints and the evolution of allometry. Evolution 68: 866–885

Ward R (1903) Records of big game: with the distribution, characteristic, dimensions, weights, and horn and tusk measurements of the different species, 4th edn. Rowland Ward, Limited, London

Westoby M, Leishman MR, Lord JM (1995) On misunderstanding the 'phylogenetic correction'. J Ecol 83:531–534

Williams GC (1992) Natural selection: domains, levels, and challenges. Oxford University Press, Oxford

Chapter 15
Modelling Stabilizing Selection: The Attraction of Ornstein–Uhlenbeck Models

Brian C. O'Meara and Jeremy M. Beaulieu

Abstract Ornstein–Uhlenbeck models are a generalization of Brownian motion models that allow trait values to evolve to follow optima. They have become broadly popular in evolutionary studies due to their ability to better fit empirical data as well as for the biological conclusions which can be drawn based on their parameter estimates, especially optimum trait values. We include a survey of available software implementing these models in phylogenetics as well as cautions regarding the use of this software.

15.1 Introduction

The mean value of a trait in a species is affected by multiple factors: physical constraints on evolution, lack of variation, change due to finite population size, and trade-offs between different optima. From one generation to the next, a trait value could change due to processes such as genetic drift, selection towards an optimum, or mutational pressure. If these movements are independent and identically distributed and have an additive effect through time, by the central limit theorem, evolution will fit a Brownian motion process (if the movements have a multiplicative effect through time, the log of the trait value will be evolving under Brownian motion). An Ornstein–Uhlenbeck (OU) process would better describe the process if these movements tended to be in the direction of a particular trait value (such that species with a trait value larger tend to evolve a smaller trait value).

B. C. O'Meara (✉)
Department of Ecology and Evolutionary Biology, Knoxville, TN, USA
e-mail: bomeara@utk.edu

J. M. Beaulieu
National Institute for Biological and Mathematical Synthesis, University of Tennessee, Knoxville, TN 37996, USA

By way of a rough example, consider the position of a toddler attached to their parent by an elastic band. On average, the toddler's position is centred on the parent's position. This mean trait position (with trait units) is often denoted θ and called the "optimum" as an analogy to models of adaptive quantitative evolution. In some phylogenetics models, θ is fixed; in others, there can be discrete shifts in optima (equivalent to attaching the parent end of the lead to a different parent), and in others, θ can move (equivalent to a parent walking down the street with toddler attached). The rate at which a toddler wiggles, σ^2 (in units of squared trait value over time), corresponds to the equivalent parameter with Brownian motion. Finally, the strength of pull by the band (α) can also vary: a strong band pulls the toddler back to the parent more rapidly than a weak band does. A different way to express this is in terms of the amount of time that is expected for a trait to move halfway to the mean value (phylogenetic half-life) which is simply $\ln(2)/\alpha$ (and is measured in time units) (Hansen 1997). For more detailed information about OU models, see Chap. 14.

15.2 Utility

In phylogenetics, we can use the OU model to describe the motion of one trait that depends on the state of another trait or the motion of a trait thought to be constrained. Note that in the former case, the independent trait need not be explicitly included in the model. OU models are often interpreted as models of adaptation, with θ thought to be an adaptive optimum and σ^2 thought to be variance due to genetic drift. This is generally an incorrect interpretation, however. As noted by Hansen (1997), on a macroevolutionary timescale, there is an almost instantaneous movement of a trait to its optimum, in contrast to the half-life of millions of years often discovered from application of OU models. Such a model, therefore, describes how a trait optimum itself moves. If a trait value for a species is not at an optimum, it is more likely that the true optimum of that species is not at the θ given in the model rather than that the trait value is itself far from the optimum.

On the whole, OU models in phylogenetics are simply phenomenological models of optimum movement rather than quantitative genetics models of adaptation within species. Nonetheless, these models have substantial utility. First, they can adjust for nonlinear accumulation of variance with time. Under Brownian motion, two diverging species have trait variance that increases linearly with time: species sharing an ancestor 50 MYA have twice the trait variance of species sharing an ancestor 25 MYA. However, on a long timescale, we might expect this rate to slow down: two sister species of flowering plants may differ in height by 10 % after just a short divergence time, but we do not expect over many millions of years to have one species microscopic and the other taller than the tallest redwood. This pattern of rate of trait divergence slowing through time (due to factors such as soft constraints on trait values) can be fit by using a single mean OU model, which has the effect of shortening especially rootward edges (in units

of expected change) in comparison with the expectation under Brownian motion. This correction can be useful for models such as independent contrasts (Felsenstein 1985) that rely on branch lengths in units of expected change.

These models can also be used to test evolutionary hypotheses. For example, Whittall and Hodges (2007) evaluated the idea that there are three distinct evolutionary optima for nectar spur length based on pollinator type using OU models. They mapped on OU regimes based on pollinator syndrome and compared the fit of an OU model with three θ parameters (one parameter per regime) with fits of models that had just one optimum or Brownian motion. Many studies use a similar strategy of comparing models with multiple pre-assigned optimum trait value parameters with models with a single optimum or a continually moving optimum (Brownian motion). Recent models (Beaulieu et al. 2012) allow for θ, σ^2, and/or α to all vary on the tree. For example, one could investigate whether the rate of evolution, σ^2, varies over the tree but use a single optimum. Figure 15.1 shows simulations of these models.

Another use of these models is in a more exploratory vein. Ingram and Mahler (2013) developed an approach that rather than a priori assignment of θ parameters to regimes on the tree allows the data and tree to drive this assignment (see also Chap. 18). This allows detection of unexpected clumping of optimality parameters. However, this can be interpreted in a hypothesis-testing framework as well. For example, Mahler et al. (2013) examined *Anolis* lizards and find support for convergence of morphological optima, consistent with earlier work on this group (e.g. Losos 1992; Jackman et al. 1997), but also recovered unique optima that could be investigated in the future.

Regardless of whether a model is being used to test a hypothesis or to investigate parameters, one must choose which model to use. Sometimes, this is the actual question: Is a model with regimes mapped based on habitat better than a model with regimes mapped based on diet? In other cases, the relevant question is the actual parameter estimate: Is the optimal body size for mammals in the temperate region greater than the optimal size for tropical mammals? For an investigation where the model chosen is of primary interest, a likelihood ratio test (comparing the fit of two nested models) or a reversible jump MCMC approach, where an algorithm can move between different models, returning the posterior probability of each, would be appropriate. Many biologists incorrectly use the Akaike information criterion as a proxy for significance, though it is appropriate for determining which model loses the least amount of information as well as relative weights for different models. Biology is complex: Ornstein–Uhlenbeck models describe a phenomenological process, and it is unlikely that all taxa in a clade have exactly the same OU parameters. Given enough power, a more complex model is likely to be chosen. For model comparison questions, therefore, it is important to consider carefully what the actual question is, as ability to reject a simple BM or OU model may come more from power than a biological process. For many questions, we recommend a parameter estimation approach instead, rather than merely demonstrating that Brownian motion is rejected in favour of a single mean OU model, which shows that under the best model, phylogenetic half-life is greater than the age of the tree, suggesting a very slow pull towards an
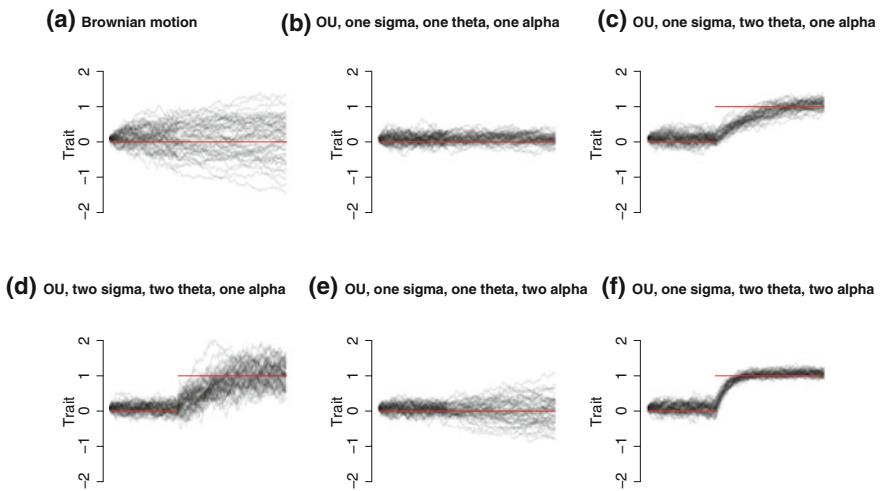
Fig. 15.1 Simulations of OU processes. On all plots, a *red line* shows the θ value for that time period, and 50 independent simulations are shown. **a** shows Brownian motion, with the same σ^2 as used in the other plots. **b** has the same parameters as **a**, except that now α is positive, making it an OU rather than BM process. **c** shows a shift in θ part-way through the simulation. Note how the approach of the simulations to θ slows as they get closer. **d** is the same as **c**, but with σ^2 increasing at the same time as the θ shifts. **e** matches **b**, except that α decreases part-way through the simulation, allowing more deviation from the optimum value. **f** matches **c**, but with α increasing at the same time as the θ shifts. **a** Brownian motion, **b** OU, one sigma, one theta, one alpha, **c** OU, one sigma,two theta, one alpha, **d** OU, two sigma,two theta, one alpha, **e** OU, one sigma,one theta, two alpha, **f** OU, one sigma,two theta, two alpha

optimal value. Biological significance of parameters should be considered as well as statistical significance. Given frequent uncertainty in choosing the best model, a multimodel inference approach (Burnham and Anderson 2004) may be appropriate where inferences about parameter values are based on multiple models weighted by their fit (often using Akaike weights) rather than just from a single model.

15.3 Historical Development

Ornstein–Uhlenbeck models have a long history in physics (Uhlenbeck and Ornstein 1930; Doob 1942) and have also been used in finance (Barndorff-Nielsen and Shephard 2001). Within phylogenetics, their adoption was proposed by Felsenstein (1988), yet their widespread use can be traced to their advocacy by Hansen (1997) and development in an information-theoretic context by Butler and King (2004). In fact, Butler and King's (2004) R package OUCH was the first to provide biologists a useful framework for analysing models of Brownian motion and Ornstein–Uhlenbeck models with one or more means. With OUCH, the means are painted as "regimes" on the tree by assigning different evolutionary parameters to different

branches of a tree. For example, one could imagine that all parts of the tree that are reconstructed as having bats as pollinators may be assigned to have θ_1, while the parts of the tree reconstructed as having insect pollinators could be assigned θ_2. One could also assign regimes based on taxonomy: assign one θ parameter to the angiosperm clade and a different θ parameter to the paraphyletic group of non-angiosperm vascular plants. The optimal values for these parameters can be found under likelihood. Models (differing in number and mapping of regimes, as well as non-OU models) may be compared using the Akaike information criterion (Akaike 1973). Chapter 12 covers model selection in more detail. In this way, a model with one mean parameter can be compared with one with multiple mean parameters. Beaulieu et al. (2012) extended this by allowing for θ, σ^2, and/or α to all vary on the tree, in their R package OUwie. This also allows painting of sets of model parameters (regimes) on different parts of the tree. In theory, regimes may change within a branch, but by default, many programs assume no more than one change per branch and assign regimes to nodes. In OUwie, if two ends of a branch differ in regime, the regime change is assumed to happen halfway along a branch; in OUCH, in this situation, the regime change is assumed to happen at the beginning of the branch. In OUwie, stochastic character mapping (Huelsenbeck et al. 2003) from the R package phytools (Revell 2012) can be used to reconstruct the state of a discrete character everywhere along a tree and then assign regime based on this reconstruction. However, this can cause issues if there is uncertainty in this mapping (Revell 2013), as some parts of branches will be misassigned to the wrong regime.

A separate trend has been the development of multivariate models. Hansen et al. (2008) developed an approach to relate a character evolving under an OU process to a mean evolving under a BM model. This allows for a natural lag between the state of the predictor variable (the trait evolving under Brownian motion) and the state of the character evolving under OU (which has an optimum value that depends on the state of the predictor variable). This was later extended by Bartoszek et al. (2012) to a case where multiple characters are coevolving, perhaps in addition to a predictor variable.

A quasi-multivariate approach was developed by Ingram and Mahler (2013) in the R package SURFACE. This wraps OUCH, so uses its model with a single σ^2 and α value for a character and one or more θ values. However, it tries different painting of θ regimes over the tree and finds the painting that minimizes the information lost in the model. Its quasi-multivariate nature is that the painting of the regimes, but not the parameter values themselves, is shared across multiple characters, allowing them to jointly inform this placement.

Finally, an OU model with a single regime everywhere on the tree has the effect of simply transforming branch lengths, in the same way that Pagel's kappa or lambda (Pagel 1997, 1999) are ways to stretch a tree so that branch lengths better represent the evolutionary process. Thus, there are several software packages that can choose a single α value that results in a set of branch lengths that best fit a model for available data, either for a single trait or for dealing with phylogenetic non-independence when doing regression or correlation between multiple traits. These include PHYSIG (Blomberg et al. 2003), ape (Paradis et al. 2004),

Table 15.1 Models implementing OU processes for trait evolution

Software name	Citation	Maximum number of regimes	Allows variation in θ parameters	Allows variation in σ^2 parameters	Allows variation in α parameters	Regression	Ancestral state estimation at multiple nodes	Software type
PHYSIG	Blomberg et al. (2003)	One per tree	N	N	N	Y	N	MATLAB scripts
OUCH	Butler and King (2004)	One per branch	Y	N	N	N	N	R package
Ape	Paradis et al. (2004)	One per tree	N	N	N	N	Y	R package
COMPARE	Martins (2004)	θ varies with other traits	Y	N	N	Y	Y	Java program
Geiger	Harmon et al. (2008)	One per tree	N	N	N	N	Y	R package
SLOUCH	Hansen et al. (2008)	θ varies with other traits	Y	N	N	Y	N	R scripts
OUwie	Beaulieu et al. (2012)	Arbitrarily high	Y	Y	Y	N	N	R package
MvSLOUCH	Bartoszek et al. (2012)	θ varies with other traits	Y	N	N	Y	N	R package
SURFACE	Ingram and Mahler (2013)	One per branch	Y	N	N	N	N	R package
Phylolm	Ho and Ané (2014)	One per tree	N	N	N	Y	N	R package

COMPARE (Martins 2004), geiger (Harmon et al. 2008), and phylolm (Ho and Ané 2014). Table 15.1 compares multiple software packages which implement OU models for phylogenetics.

15.4 Caveats

With Brownian motion, the expected value after any amount of time is the initial value. Even so, as time information about ancestral state decays, so does information about potentially different models operating deeper in the tree. With Ornstein–Uhlenbeck processes, the expected value is a weighted average of the initial value and the optimal value. Longer amounts of time, and stronger attraction parameters, mean that historical signal will begin to disappear. Even for large trees, there may be very little information about past regimes. Uncertainty in returned parameter values may be estimated by using an approximation based on the slope of the likelihood surface at its point of maximum likelihood (Beaulieu et al. 2012). A better way to estimate this is to look at the actual likelihood surface over a range of parameter values (as in SLOUCH (Hansen et al. 2008)). Parametric bootstrapping (simulating under the recovered model) is another way to estimate uncertainty: Under the assumption that a model is true, what distributions of parameter estimates are recovered if evolution were rerun under that model? High α values erase history about past processes, but low α values may also be problematic in that they make multiple θ parameters more difficult to estimate, as the final trait values could depend less on θ. In addition to issues arising with a low or high α, some of the more complex models with multiple α and σ^2 values can also be difficult for parameter estimation. The number of regimes on a tree can increase without limit, as each branch can be broken into multiple regimes, and this can rapidly exhaust any information in the data. Even assuming no more than one regime per branch and thus a branch-specific estimate of α, σ^2, and θ, on a tree with N taxa, there are $3 \times (2N-2)$ parameters to estimate but no more than $N \times$ number of characters (typically one) to provide data.

Dealing with the state at the root can also be problematic. With Brownian motion, the root state is estimated based on the parameter values and branches and does not depend on the rate of evolution. With Ornstein–Uhlenbeck processes, the root state can be estimated, but where there is little information about the past due to strong attraction, it is biased towards zero (Beaulieu et al. 2012). Some software, such as OUwie or phylolm, default to assuming the root state comes from the stationary distribution of the evolutionary process, but has an option to separately estimate the root state. However, the lack of information about the root state can mean that separately estimating this parameter can lead to inaccurate estimates for other parameters. It is important to note that while OUCH originally estimated a state at the root, more recent versions of the package assume stationarity. Bartoszek et al. (2012) fix the root state at the optimal value for the regime on the root branches.

A related issue is whether Brownian motion is a restriction of an Ornstein–Uhlenbeck model. In other words, as α approaches zero, do the model's parameter estimates and likelihood converge towards those of a Brownian motion model? There is, of course, the caveat that the uncertainty in estimates of θ should increase without bound as one approaches α of zero. Nevertheless, since the treatment of the root state differs between OU model implementations, but not between Brownian motion implementations, in some software (such as OUwie or geiger), an OU model with one regime approaches Brownian motion as α approaches zero. However, in others, such as the current version of OUCH, the implemented Brownian motion model is not nested within the implemented OU model: the likelihood of an OU model with a single θ does not converge to the likelihood of a Brownian motion model as α approaches zero. This is an active area of discussion that has yet to be resolved.

An important issue when dealing with any model is understanding what the parameters mean. The θ parameter is in the units of the trait(s) under investigation (i.e., kg for body mass), and this is true regardless of implementation. The σ^2 rate is in units of trait units squared over branch length units (i.e., kg/MY), and α is in units of reciprocal branch length units (i.e., MY^{-1}). Phylogenetic half-life has time units (i.e., MY). One unfortunate trend, in both our work and the work of others, is to fail to report these units. Users should note that some programs in this area rescale trees before analysis. One common rescaling is dividing each branch length by the total height of the tree, which makes all root to tip lengths one. This can help deal with numerical issues arising from software with finite precision, but it also means that σ^2 and α will not have the same units as in the original tree, hindering interpretation. However, these parameters can be rescaled to the original units later in the process.

Biologists have become used to computational issues involved in tree inference and so treat such inferences with caution; fitting a single model to an existing tree can seem like a trivial issue in comparison. However, estimating the likelihood of an OU model and getting good parameter estimates for it can, in practice, be difficult. Part of this stems from nearly flat likelihood surfaces: numerical optimization may terminate before reaching a peak if changes in parameter values have only a slight effect on the likelihood. Certain pairs of parameters may also tend to form a ridge in likelihood space, making them difficult to optimize. There is also the temptation to use overly complex models: a thirty-taxon dataset may simply not have enough information in one character to estimate multiple α values. While model selection approaches should not choose models that are too complex for the available data, they are not guaranteed to work all the time. The details of the mathematical steps used to calculate likelihood for OU models in many software packages can make these methods prone to encountering errors due to finite precision, though these errors may be hidden from most users (but may still affect whether optimization works properly). Many of the developers of software packages are biologists or mathematicians foremost who may lack the expertise to test and optimize every part of a program. For these various reasons, users of OU software should examine results with some skepticism. The open source nature of

software in this field will allow users to examine its interior workings and do things such as try different starting points for parameter optimization (some programs do this automatically, others do not). However, even more basic tests can be performed without knowing how to program that will give a sense of whether the parameter estimates coming back are reasonable or may reflect a problem with the software. For example, if the exact same analysis is run again, are the same likelihood scores and parameter estimates returned each time? Inconsistency may indicate that the search is sensitive to starting values and that more runs must be attempted to find the best values. While not all models compared need to be nested, in cases where one model is truly a restriction of another model, the likelihood of the restricted model (but not the AIC score) must be the same or worse than the likelihood of the general model: Does this occur? For programs that allow user specification of fixed points at which to evaluate likelihood, what is the shape of the likelihood surface: Are points near the returned maximum likelihood estimates of the parameters always worse in likelihood than what the program returned as the best values? In the case where two programs implement the same model, do they return the same parameter estimates for the same data (though note the issue about potential rescaling, above)? Is a returned parameter value at one of the preset maximum or minimum bounds of the software? If so, it may make sense to change this bound, as the maximum likelihood estimate is probably outside this region, though this may result in numerical precision problems within the program (often a reason for setting default bounds). Doing these steps takes time, but spending a few extra days to verify that the returned results are correct may be a worthwhile investment after the months to years it can take to get the phylogeny and trait data required to address a biological question using these models.

15.5 Example

Within flowering plants, there is a strong growth form-dependent distribution in genome size (i.e., the amount of DNA in any given cell), with woody species containing smaller genome sizes, on average, as well as lower overall variance, when compared to herbaceous species. It has been suggested (Beaulieu et al. 2008, 2010, 2012) that these patterns largely reflect differences in life history. Woody angiosperms generally take longer to reach reproductive maturity (Verdú 2002), leading to longer generation times and fewer opportunities for random insertion/ deletions to occur on a per unit time basis. Indeed, woody lineages consistently show slower overall rates of genome size evolution when compared to herbaceous lineages (Beaulieu et al. 2010, 2008).

Here, we illustrate how the OU framework can be used to uncover rate differences in genome size evolution between woody and herbaceous growth form states within the Fabaceae (i.e., legumes). This example is further fleshed out in the Online Practical Material (hereafter OPM) available at http://www. mpcm-evolution.com. Genome size estimates were taken from the Plant DNA

C-value database (Bennett and Leitch 2010), and we arrayed onto the time-calibrated phylogeny of legumes from Beaulieu et al. (2010). We focus these analyses on the monoploid genome size, or the 1Cx value, in order to correct for the possibility that polyploidy can inflate rates of genome size evolution. The monoploid genome size represents the amount of DNA in the unreplicated monoploid chromosome set and is calculated by dividing the 2C DNA amount by ploidy. These values were $\log_{10}$-transformed prior to all analyses to ensure that these data minimally conformed to Brownian motion evolution (Oliver et al. 2007).

For this analysis, regimes are mapped on the tree based on likelihood estimation of ancestral states using the package corHMM (Beaulieu et al. 2013) though stochastic character mapping, parsimony, or other ways of assigning regimes to branches could be used. A variety of continuous trait models are then investigated. These include Brownian motion with a single rate ("BM1" in the program), Brownian motion with a different rate allowed for each discrete state regime ("BMS"), OU with a single optimum ("OU1"), OU with a different optimum for each regime, but with a constant α and rate of evolution ("OUM"), and finally a model with a different optimum and rate of evolution for each regime ("OUMV"). Other models are available, such as one that varies optimum, α, rate of evolution for each regime or various other restrictions of this model, but the set of models for the example is limited to a workable set which can run relatively quickly.

For all the models, the parameter estimates as well as the AIC with small sample correction (AICc) are stored. The best (smallest) AICc value is subtracted from the AICc values for each model to get the ΔAICc value. In this example, the best model is OUMV, a model with one α across the tree but different σ^2 and θ for each discrete state regime. The next best model has a ΔAICc much higher (42.8), which corresponds to an Akaike weight that is much tinier (<1 billionth) than that of the best model. If support were much more similar across models, model averaging would make sense as a way to deal with uncertainty in the models, but given this difference is would not be expected to help. In fact, model averaging may even be problematic if a complex model with very little support generates very bad parameter estimates, as the low weight on the model might not be enough to counter the magnitude of the poor estimates. For the best model, the optimal value for herbaceous plants was $0.254 \pm 0.036 \log_{10}$ (pg), which corresponds to 1.289 ± 1.037 pg, while for woody plants, it was 0.904 ± 1.029 pg after transformation out of log-space. The half-life of the process is 0.115 MY, while the tree is 59 MY old, suggesting a strong pull to each of these values. The rate of evolution of genome size while herbaceous is over five times greater than the rate of evolution while woody ($2.968 \pm 0.303 \log_{10}$(pg)/MY vs. $0.574 \pm 0.105 \log_{10}$ (pg)/MY).

There are two lessons from this analysis. The first is that history does not matter much for genome size in this group: the short half-life indicates that the trait value of a species is quickly pulled from its ancestral state to whatever the optimal state is. However, the optimal states seem very similar, and each is within one standard error of the other, suggesting little evidence for different optimal states. In contrast, the σ^2 rates do differ, having a biologically significant fivefold rate difference as well as a statistically meaningful rate difference. This suggests that while history

does not matter much for mean value, state matters a great deal for rate of evolution. This analysis also points out a limitation of software: given the similarity in OU means across states but not OU σ^2 parameters, an even better model might be one with one α and θ across the tree but different σ^2 for each discrete state (in OUwie's jargon, this would be an OUV model) but this has yet to be implemented.

15.6 Future Directions

There are numerous potential advances in this area. One trivial advance would be inference of ancestral states under an OU process. This is possible with a single OU mean tree transform, as can be implemented in geiger (Harmon et al. 2008) or COMPARE (Martins 2004), but has not yet been implemented in software for more complex OU models. Another straightforward advance would be a wrapping of the Beaulieu et al. (2012) family of OU models in the SURFACE (Ingram and Mahler 2013) approach to check for regime shifts for σ^2 and α in the same way this is done for OUCH-type models.

Most work in this area has been in a regression or information-theoretic framework, and the utility of Bayesian approaches has yet to be explored fully. They have potential in allowing a way to bring in information from external sources about parameters as priors without having to fix this information. However, given frequent uncertainty in parameter estimates in these models, it will be essential to make sure that the results are driven in part by the data rather than only reflecting the priors.

Ornstein–Uhlenbeck models are among the most complex models of continuous trait evolution available to date. They give information about the parameters of evolutionary change on a macroevolutionary timescale but may not reflect microevolutionary processes (Hansen 1997). Models that operate at the level of population genetics mechanisms may be important in the future to allow inferences of processes rather than just fitting evolutionary patterns. Continued development of multivariate approaches remains important as well.

References

Akaike H (1973) Information theory as an extension of the maximum likelihood principle. In: Petrov BN, Csaki F (eds) Second international symposium on information theory. Akademiai Kiado, Budapest, pp 267–281

Barndorff-Nielsen OE, Shephard N (2001) Non-gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics. J Roy Stat Soc Ser B (Stat Methodol) 63(2):167–241. doi:10.2307/2680596

Bartoszek K, Pienaar J, Mostad P, Andersson S, Hansen TF (2012) A phylogenetic comparative method for studying multivariate adaptation. J Theor Biol 314:204–215

Beaulieu JM, Jhwueng D-C, Boettiger C, O'Meara BC (2012) Modeling stabilizing selection: expanding the Ornstein–Uhlenbeck model of adaptive evolution. Evolution 66(8):2369–2383

Beaulieu JM, Leitch IJ, Patel S, Pendharkar A, Knight CA (2008) Genome size is a strong predictor of cell size and stomatal density in angiosperms. New Phytol 179(4):975–986

Beaulieu JM, O'Meara BC, Donoghue MJ (2013) Identifying hidden rate changes in the evolution of a binary morphological character: the evolution of plant habit in campanulid angiosperms. Syst Biol 62:725–737

Beaulieu JM, Smith SA, Leitch IJ (2010) On the tempo of genome size evolution in angiosperms. J Bot. doi: 10.1155/2010/989152

Bennett MD, Leitch IJ (2010) Plant DNA C-values database (release 6.0, Dec. 2012)

Blomberg SP, Garland T, Ives AR (2003) Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution 57(4):717–745

Burnham KP, Anderson DR (2004) Multimodel inference—understanding AIC and BIC in model selection. Sociol Methods Res 33(2):261–304

Butler MA, King AA (2004) Phylogenetic comparative analysis: a modeling approach for adaptive evolution. Am Nat 164(6):683–695

Doob JL (1942) The Brownian movement and stochastic equations. Ann Math 43(2):351–369

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

Felsenstein J (1988) Phylogenies and quantitative characters. Annu Rev Ecol Syst 19:445–471

Hansen TF (1997) Stabilizing selection and the comparative analysis of adaptation. Evolution 51(5):1341–1351

Hansen TF, Pienaar J, Orzack SH (2008) A comparative method for studying adaptation to a randomly evolving environment. Evolution 62(8):1965–1977

Harmon LJ, Weir JT, Brock CD, Glor RE, Challenger W (2008) GEIGER: investigating evolutionary radiations. Bioinformatics 24(1):129–131

Ho LST, Ané C (2014) A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. Syst Biol 63(3): 397–408

Huelsenbeck JP, Nielsen R, Bollback JP (2003) Stochastic mapping of morphological characters. Syst Biol 52(2):131–158. doi:10.1080/10635150390192780

Ingram T, Mahler DL (2013) SURFACE: detecting convergent evolution from comparative data by fitting Ornstein–Uhlenbeck models with stepwise Akaike Information Criterion. Methods Ecol Evol 4(5):416–425

Jackman T, Losos JB, Larson A, de Queiroz K (1997) Phylogenetic studies of convergent adaptive radiations in Caribbean Anolis lizards. In: Molecular evolution and adaptive radiation. pp 535–557

Losos JB (1992) The evolution of convergent structure in Caribbean Anolis communities. Syst Biol 41(4):403–420

Mahler DL, Ingram T, Revell LJ, Losos JB (2013) Exceptional convergence on the macroevolutionary landscape in island lizard radiations. Science 341(6143):292–295

Martins EP (2004) COMPARE, version 4.6 b. Computer programs for the statistical analysis of comparative data. Distributed by the author at http://compare.bio.indiana.edu/, Department of Biology, Indiana University, Bloomington, IN

Oliver MJ, Petrov D, Ackerly D, Falkowski P, Schofield OM (2007) The mode and tempo of genome size evolution in eukaryotes. Genome Res 17(5):594–601

Pagel M (1997) Inferring evolutionary processes from phylogenies. Zoolog Scr 26(4):331–348

Pagel M (1999) Inferring the historical patterns of biological evolution. Nature 401(6756):877–884

Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20(2):289

Revell LJ (2012) Phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol Evol 3(2):217–223

Revell LJ (2013) A comment on the use of stochastic character maps to estimate evolutionary rate variation in a continuously valued trait. Syst Biol 62(2):339–345

Uhlenbeck GE, Ornstein LS (1930) On the theory of the Brownian motion. Phys Rev 36(5):823–841

Verdú M (2002) Age at maturity and diversification in woody angiosperms. Evolution 56(7):1352–1361. doi:10.1111/j.0014-3820.2002.tb01449.x

Whittall JB, Hodges SA (2007) Pollinator shifts drive increasingly long nectar spurs in columbine flowers. Nature 447(7145):706–709

Chapter 16
Hidden Markov Models for Studying the Evolution of Binary Morphological Characters

Jeremy M. Beaulieu and Brian C. O'Meara

Abstract Biologists now have the capability of building large phylogenetic trees consisting of tens of thousands of species, from which important comparative questions can be addressed. However, to the extent that biologists have applied these large trees to comparative data, it is clear that current methods, such as those that deal with the evolution of binary morphological characters, make unrealistic assumptions about how these characters are modeled. As phylogenies increase both in size and scope, it is likely that the lability of a binary character will differ significantly among lineages. In this chapter, we describe how a new generalized model, which we refer to as the "hidden rates model" (HRM), can be used to identify different rates of evolution in a discrete binary character along different branches of a phylogeny. The HRM is part of a class of models that are more broadly known as Hidden Markov models because it presupposes that unobserved "hidden" rate classes underlie each observed state and that each rate class represents potentially different transition rates to and from these observed states. As we discuss, the recognition and accommodation of this heterogeneity can provide a robust picture of binary character evolution.

16.1 Introduction

Underlying many important discoveries in ecology, evolution, and behavior is the use of a phylogenetic tree. Phylogenies allow for the non-independence of taxa to be accounted for while also opening up new ways of examining how traits change

J. M. Beaulieu (✉)
National Institute for Biological and Mathematical Synthesis,
University of Tennessee, Knoxville, TN 37996, USA
e-mail: jbeaulieu@nimbios.org

B. C. O'Meara
Department of Ecology and Evolutionary Biology,
University of Tennessee, Knoxville, TN 37996, USA
e-mail: bomeara@utk.edu

through time, detect correlations between ecological and morphological characters, and better understand patterns of lineage diversification. Prior studies of these questions, particularly of larger, older, and widespread clades, have tended to rely on very sparse taxon sampling or on a representative sample of the major lineages contained within the group. The traits amenable to these types of trees can neither change too little (lest there be no variation to examine) nor too much (lest there be no signal left to detect) and so are naturally a biased set. The amount of sequence data available has grown rapidly. In just 20 years, what has been considered a "large" tree has gone from 500 taxa (Chase et al. 1993) to those that contain more than 50,000 species (Smith et al. 2011, also see Bininda-Emonds Chap. 4 this volume). By scaling up analyses to include many more taxa, we can analyze traits with a much wider range of evolutionary rates, and this includes many traits of great ecological and evolutionary importance.

At the same time, however, it is clear that these large comprehensive phylogenies present new challenges for comparative biology. For decades, the dominant models have always assumed a homogeneous process through time and across taxa. As biologists, we know that life is not evolving according to a homogeneous process. Mass extinction events change the ecological context for evolution. Evolution within lineages can lead to different selective regimes. Such heterogeneity requires that different models be applied to different parts of a phylogeny. Recent attention has been focused almost exclusively on solving this problem for continuously varying characters. It is now possible to apply parameter-rich models for detecting meaningful differences in phenotypic evolution among clades, among specific branches, or even pieces of branches under processes such as Brownian motion (O'Meara et al. 2006; Thomas et al. 2006; Revell 2008) or the Ornstein–Uhlenbeck process (Butler and King 2004; Beaulieu et al. 2012). Models of discrete binary character evolution, on the other hand, have received fairly little attention, and biologists are still forced to rely on conventional models that apply uniform rates of change to all branches in a tree.

Simple models of binary character evolution may make sense for less inclusive clades, such as the traditional genus or family levels, which often contain relatively few instances of character change. But, they are not likely to adequately explain the evolution of a discrete binary character in very large and very old clades. At these phylogenetic scales, it is hard to ignore evident distributions of observed character states. Some parts of the tree will often only exhibit one character state, while other parts apparently have undergone frequent state changes and appear rather labile. Obviously, not accounting for heterogeneity in different parts of a tree can lead to problems with estimates of transition rates and/or the inference of the likeliest ancestral states that accompany them. More broadly, however, when phylogenetic trees can contain many thousands of species, we miss important opportunities to discover patterns that could not previously have been recognized and quantified.

In this chapter, we describe a generalized model, which we refer to as the "hidden rates model" (HRM) (Beaulieu et al. 2013a) that allows for the identification of different rates of evolution in a discrete binary character along different

branches of a phylogeny. This model was inspired by the covarion model (COncomitant VARIable codON; Fitch and Markowitz 1970) of nucleotide substitution, which presupposes that unobserved rate classes underlie each nucleotide state at a site in an alignment, and that each rate class represents potentially different transitions to and from observed states (Penny et al. 2001; Galtier 2001). All these models are part of a class that are more broadly known as Hidden Markov models because the different rate classes are treated as "hidden" states in the Markov process. As we will discuss, our HRM provides a powerful tool for detecting various forms of branch-specific heterogeneity when it exists, and can adequately pinpoint where underlying, but unobserved or unmeasured, factors have influenced the evolution of a binary character, even though the HRM is formally time homogeneous.

16.2   The Hidden Rates Model

Although they emphasize Bayesian methodology, Currie and Meade (Chap. 13 this volume) provide a thorough introduction to the underlying theory of continuous-time Markov models and the various procedures used for calculating their likelihood and estimating transition rates. But briefly, under a likelihood-based approach, the likelihood is defined as being proportional to the probability of observing the data given a model of evolution and a specific tree,

$$L(\mathbf{Q}) \propto P(\mathbf{D}|\mathbf{Q}, \mathbf{T}) \qquad (16.1)$$

where the data, $\mathbf{D}$, is a vector of observable character states at the tips of a phylogenetic tree, $\mathbf{T}$, whose branch lengths and topology are assumed to be known. The model of evolution, defined by $\mathbf{Q}$, is an instantaneous rate matrix describing the possible transition rates between character states. For a single binary character that has two observable states, 0 and 1, $\mathbf{Q}$ is a 2×2 matrix,

$$\mathbf{Q} = \begin{bmatrix} - & q_{0 \to 1} \\ q_{1 \to 0} & - \end{bmatrix} \qquad (16.2)$$

which we can then use to compare the fit of two models: one where we assume equal transition rates between the two states ($q_{0 \to 1} = q_{1 \to 0}$), or one where we assume two distinct transition rates ($q_{0 \to 1} \neq q_{1 \to 0}$). This matrix is transformed into a transition-probability matrix, symbolized as $P(t)$, and is equal to $e^{\mathbf{Q}t}$, where t represents the length a branch. In general, we use this matrix to calculate conditional likelihoods, which are defined as the sum of the probability of observing everything descended from a focal node given that the focal node is in each character state. These likelihoods are computed for every node in the tree starting with the tips and working down toward the root. Thus, the conditional likelihood at the root represents the likelihood in Eq. (16.1) (for the more details about the dynamic programming

algorithm used to carry out this computation, see Felsenstein 1981). In order to complete the calculation at the root, however, an additional step is needed which involves of weighting the conditional likelihood by the prior probability of the possible states at the root. By default, we assume that each possible character state is weighted equally. Other approaches weight the conditional likelihood by the probability that each character state gave rise to the descendant character states given the transition rates and the tree—a procedure described by FitzJohn et al. (2009). This probability is calculated by dividing the likelihood that the root is in each character state and rate combination by the sum of the likelihoods of all possible character states.

Looking at $\mathbf{Q}$ in Eq. (16.2), it is easy to see how unsatisfying it might be to assume that, at most, two transition rates (a forward and backward rate between the two states for our character) govern the evolution of a binary character, particularly when applying such a model to a very broad assemblage of species. As one zooms out, including more and more clades, the factors that are associated with transitions between states are unlikely to be consistent. For example, the frequency of transitions between fleshy and dry fruit types in flowering plants will differ depending on whether or not clades occur in regions where biotic dispersal is more likely (i.e., tropics). In insects, the loss of flight will vary based on the environment, costs of dispersal, or any other correlated factor that can change across a tree. More broadly, processes that can also affect rates of character evolution include generation time, effective population size, the underlying genetic architecture of the trait, and/or mutation rates. However, going into the analysis, we may be unaware or even unclear what these specific factors might be—we just know that these rates could change in different portions of the tree. Thus, the HRM is designed as a means to effectively "paint" areas of a phylogeny where transitions happen frequently or infrequently due to unmeasured characters that affect the rates.

Conceptually, the HRM is a generalized form of the covarion model (Fitch and Markowitz 1970), which is used to infer phylogenies from sequence data by allowing the rate of nucleotide substitutions to not only vary by site, but also along branches. The covarion model assumes that there are two stochastic processes at a site: one for transitions between specified rate classes; and the other for transitions between character states within a given rate class. However, because only nucleotide states can be observed, these rate classes are considered "hidden" states in the model, and therefore, we have to treat each observed nucleotide at a site as an ambiguous observation of the different unobserved rate classes. In the formulation of Penny et al. (2001), the unobserved rate classes are instances when the mutation rate of a nucleotide base is either turned "on," and transitions among the four nucleotide states are possible, or turned "off," where the mutation rate is set to zero.

The unobserved rate classes under the HRM are similar to the model described by Galtier (2001), where they need not be considered "on" or "off," but can represent distinct transition models (i.e., rate class A, rate class B, etc.). In other words, these rate classes may differ in being considered "on" or "off," "fast" or

"slow," or in the direction of asymmetry of transitions between states. As with the covarion model, we assume that each observed character state is an ambiguous observation of the different unobserved rate classes, and we can use the same likelihood framework as in Eq. (16.1) to calculate the conditional probability of all ancestral nodes including the root (Felsenstein 1981). That is, we begin at the tips by summing over the probabilities that are compatible with our observed character state: for example, assuming two rate classes, A and B, the probability is set to 1 for both $0A$ and $0B$ given our observation of a tip being in state 0. We define a new model, $\mathbf{Q}$, to account for the process of transitioning between all character state and rate class pairs:

$$\mathbf{Q} = \begin{bmatrix} - & q_{0A \to 1A} & q_{0A \to 0B} & 0 \\ q_{1A \to 0A} & - & 0 & q_{1A \to 1B} \\ q_{0B \to 0A} & 0 & - & q_{0B \to 1B} \\ 0 & q_{1B \to 1A} & q_{1B \to 0B} & - \end{bmatrix} \qquad (16.3)$$

Notice in this particular case that the entries in $\mathbf{Q}$ describing dual transitions (state and rate class transitions occur simultaneously) are set to zero to force such transitions to either pass through the same state to a different rate, or to pass through a different state in the same rate (see Pagel 1994). This assumption can be relaxed, of course, by simply adding these transitions back into the model. As with the likelihood model described above, a nonlinear optimization routine is used to find estimates for the entries in $\mathbf{Q}$ that maximizes the conditional likelihood at the root.
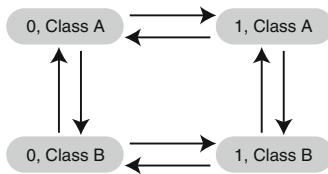
16.3 HRM Model Space

Under the HRM framework, the $\mathbf{Q}$ matrix described in Eq. (16.3) can easily be modified to allow any number of hidden rate classes (Fig. 16.1). In fact, model complexity can range from a model with a single rate class, which is the same as the familiar time-homogeneous model in Eq. (16.2), to a model that includes an infinite number of rate classes. One extreme case, where a hidden rate class is assumed for every branch, is similar in effect to the "no-common mechanism" model, which is a parsimony equivalent (Tuffley and Steel 1997). Of course, it is unlikely that such a model would ever fit the data well, because the number of parameters would far exceed the number of data points (see Holder et al. 2010). Nevertheless, we can use model selection methods, such as Akaike's information criterion (AIC) (Akaike 1974), to obtain the model that best fits the data, or calculate the relative weight for a set of models that can be taken as information about the evolutionary process. As with Huelsenbeck et al. (2004), Pagel and Meade (2006), a reversible jump MCMC approach could also be used to get the posterior probabilities of various models.
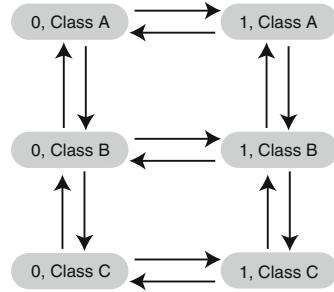
The discussion thus far has focused exclusively on evaluating models that contain increasing numbers of hidden rate classes. However, these represent a very
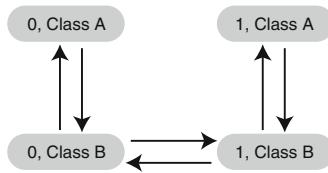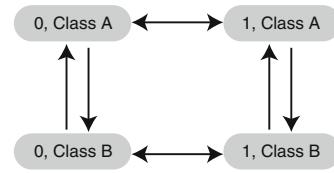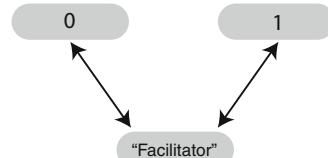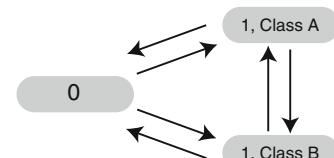
Fig. 16.1 **a** Graphical representation of HRM's with increasing number of hidden rate classes. Under the HRM framework, model complexity can range from a model with a single rate class, which is the same as a model that assumes a homogeneous process, to a model that includes an infinite number of rate classes. Here, we highlight HRM's with two- and three-ordered hidden rate classes, with the *black arrows* denoting the directions of the possible transition among the different state and rate class combinations. **b** For a given number of hidden rate classes, there are many models that contain a subset of the possible parameters (see HRM MODEL SPACE). For an HRM with two rate classes such subset models include, but are certainly not limited to, the "precursor" model of Marazzi et al. (2012), models where hidden rate classes underlie one state as opposed to both, models where particular transitions are set to zero as in the covarion model, or various combinations of models where transition rates vary among, but not within, each rate class

small subset of the possible models that could be evaluated. For instance, the "precursor model" of Marazzi et al. (2012) illustrates an efficient use of parameters in creating a specific model in the HRM framework (Fig. 16.1). Under the precursor model, there are two rates that correspond to transitions between one

observed state and a "hidden" precursor state, and between the hidden state and the other observed state. This model essentially describes the hidden state as "facilitating" transitions to and from the observed states.

To get a sense of the complete model space provided by the HRM framework, we can use Stirling numbers of the second kind (Abramowitz and Stegun 1972) to count the different combinations of models for a given a number of parameters. The Stirling numbers are computed as

$$S(n,k) = \frac{1}{k!} \sum_{i=0}^{k} (-1)^{k-i} \frac{k!}{(i-k)!i!} (i^n) \tag{16.4}$$

where n parameters are partitioned into all possible k subsets (i.e., 1, 2, …, n parameters). The Bell (1934) number, which is the sum of these Stirling numbers, can be used to count the total number of distinct model combinations contained within a given HRM. For an HRM that contains two hidden rate classes, there are 4,140 distinct models that could be evaluated (i.e., the sum of 8 total parameters partitioned into all possible subsets of 1, 2, …, 8 parameters); for three hidden rate classes there are 190, 899, 322 distinct models. Note however that this calculation assumes that the rate classes are ordered (see Fig. 16.1). This need not always be the case, and an HRM could easily be constructed that assumes an unordering of the different rate classes. Of course, the addition of several more parameters to account for transitions between all rate classes would make the model even more parameter rich and contain even more distinct subset models than in the ordered case. In any event, parameter subsets of a three-ordered hidden rate classes include, but are not limited to, models where hidden rate classes only underlie one state as opposed to both, models where particular transitions are set to zero (as in the covarion model), or various combinations of models where transition rates vary among, but not within, each rate class (Fig. 16.1). Again, we can use model selection methods to either obtain the relative weight for each of these models and average the parameters (i.e., model averaging approach), or simply determine which model is the best fit among the set.

16.4 HRM in Relation to Other Models

HRM is not the only model that deals with heterogeneity. As mentioned above, the covarion model (Penny et al. 2001; Galtier 2001) allows states to switch from off to on: effectively an extreme form of the HRM model where one of the categories has no transitions. An important set of models developed by Yang et al. (1995) fit different rates to different branches of a tree. These are typically applied for codon models but can in principle be applied to any discrete data. Yang (1994) also developed the use of a discrete gamma to deal with rate heterogeneity across sites. The likelihood of the data is calculated under each of several rates pulled from a

distribution described by a single parameter. Pagel and Meade (2004) develop a phylogenetic mixture model that is essentially a generalization of this; rather than just allowing likelihoods to be calculated across different overall rates, they allow the rate matrix to also vary.

The threshold model (Wright 1934; Felsenstein 2005, 2012) represents a different approach to dealing with heterogeneity. Rather than a hidden discrete trait affecting the rate of the observed trait, it allows for a hidden continuous trait to set the state of the observed trait. When this continuous trait, termed the liability, crosses the threshold, the discrete character changes state. This has been extended for multi-state-ordered characters (Revell 2014). The behavior of the model is grossly similar to HRM in that the frequency of discrete state changes varies over the tree; with the HRM, due to different hidden rates, and in this case, due to distance of the liability from the threshold. There are some important differences, however. With the canonical threshold model, there is the expectation over time that the liability moves away from the threshold, as it evolves with unbounded Brownian motion, and so the long-term expectation is that the transition rate eventually becomes zero. Revell (2014) has largely addressed these issues by modifying the threshold model so that the liability evolves with bounds or with a strong attraction (i.e., Ornstein–Uhlenbeck process, see O'Meara and Beaulieu Chap. 16 this volume) back to the threshold. However, the HRM also allows this in that it could have absorbing states, but it does not require it. The threshold model also assumes that near a change, where the liability is close to the threshold, the rate of gain or loss of a trait is equal, though the overall gain and loss rates could still be unequal by starting with a liability greater or lower than the threshold. The HRM can allow unequal rates on parts of the tree or over the whole tree. An advantage of the threshold model is that it only has parameters for the starting liability, the threshold value, and rate of movement of the liability; making it a relatively efficient way to fit changing rates over a tree.

16.5 Application of the HRM

The development of the HRM was motivated by the desire to understand rates of evolution between two growth habit states, woody and herbaceous, within campanulid angiosperms, a large flowering plant clade containing some 35,000 species, including the familiar composites (sunflowers and relatives), umbels (carrots and relatives), and Dipsacales (honeysuckles and relatives). Historically, growth habit has been considered by botanists to be far too labile to be understood across larger, more traditional taxonomic ranks (Cronquist 1968). This is because the vegetative features of a plant are thought to be intimately tied to the environment in which they exist, and since species contained within larger, older, and globally distributed clades can occur in a range of environments, growth habit should vary considerably at larger phylogenetic scales. Furthermore, there is increasing genetic evidence that transitions between woody and herbaceous forms should be fairly easy (Groover 2005), involving the suppression and re-expression

of only a few genes regulating both the cambial growth necessary for wood formation and the onset of flowering (Lens et al. 2012).

Such genetic integration of both vegetative and reproductive development suggests that the capacity for growth habit evolution should be ever present within flowering plants. However, it is quite puzzling that the distribution of particular habits is non-uniform across the angiosperm phylogeny. For example, some large and old clades contain only woody species (e.g., Fagales, which contain oaks and their relatives), others contain only herbaceous species (e.g., Brassicaceae, which contain mustards and their relative), and still others show extreme variation in habit presumably through many transitions between woody and herbaceous states (e.g., Asteraceae, sunflowers, and their relatives). These observations alone call into question the use of conventional likelihood-based methods for understanding the evolution of growth habit.

Here, we show results from a recent study of 8,911 campanulid species where growth habit was scored and where a large comprehensive phylogeny was used (Beaulieu et al. 2013a). We compare the fit of five models of evolution. The two simplest models are the conventional time-homogeneous models, where we either assume there are equal transition rates between woody and herbaceous states, or there are two distinct transition rates, one for transitions from woody to herbaceous and another for transitioning from herbaceous to woody. We also assessed the fit of HRM's that assume two, three, and four hidden rate classes underlying each observed woody and herbaceous state (see Fig. 16.1 for how these models are graphically structured).

When comparing the fit of the time-homogenous models to the HRM's with different numbers of hidden rate classes, it is clear that models of branch-specific rates of evolution are by far the better fit to the growth habit data. For instance, the addition of just a single hidden category provides an extraordinary improvement in the likelihood over both the one-rate and two-rate time-homogeneous models (just over 400 log-likelihood units; Fig. 16.2). With three hidden rate classes the likelihood is improved by an additional 100 log units, but begins to plateau, where the fit of a model with four hidden rate classes did not substantially improve the likelihood any further.

A HRM with three rate classes was the best-fit model overall based on AIC and with the estimated parameters indicating a complicated process of transitioning between woody and herbaceous states in an ancestrally woody clade. Incidentally, the model suggests that the three hidden rate classes represent transition models of increasing rate: "slow," "medium," and "fast." In practice, such structured rate classes will not always result, and describing rate classes in such terms as "slow" or "fast" will be inappropriate, though it works in this particular case. In the slow-rate class, the asymmetry in transition rates is similar to the time-homogenous model in that it is more likely for herbaceous species to re-evolve a woody habit than the reverse ($q_{WS \to HS} = 0.0000$, s.e. $= \pm 0.0002$; $q_{HS \to WS} = 0.0009$; s.e. $= \pm 0.0007$). In fact, being woody in the slow-rate class represents an absorbing state: herbaceous may not evolve again once a lineage transitions into this state and rate combination. However, it is important to emphasize that rates
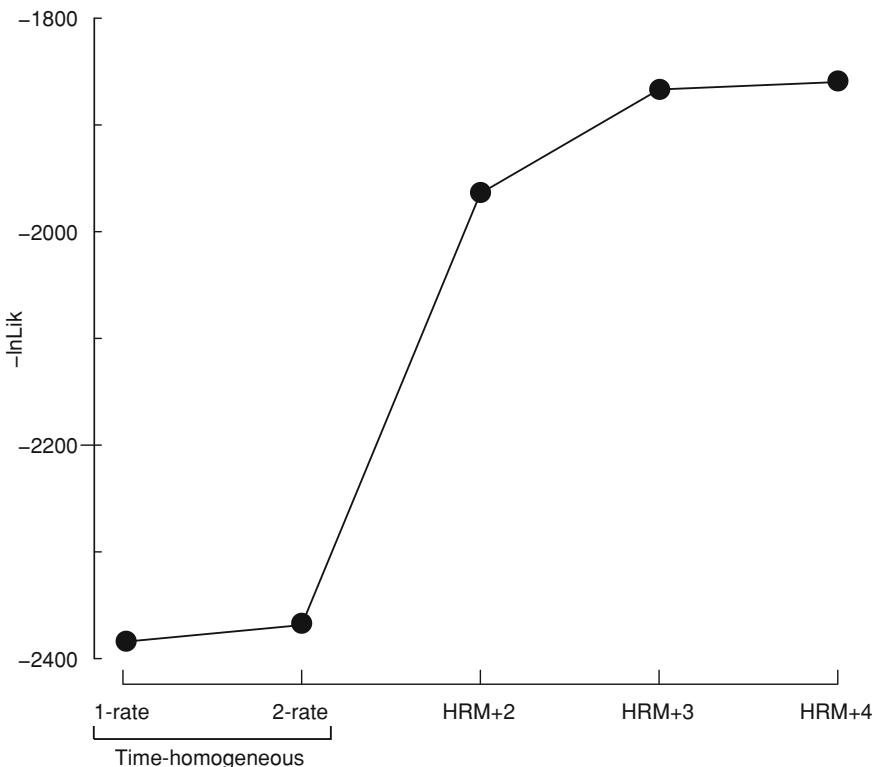
Fig. 16.2 Plot of the log-likelihoods for the different models fit to the growth habit data of Campanulidae (campanulids) from Beaulieu et al. (2013a, b). The addition of a single hidden rate (HRM + 2) improved the likelihood by just over 400 log units over both time-homogenous models. When three rate classes were allowed (HRM + 3), the likelihood was improved by another 100 log units. Four rate classes (HRM + 4) did not substantially improve the likelihood any further

under the HRM are equivalent to substitution rates, not mutation rates, and therefore, only provide an indication of what happened over evolutionary time and not what was proposed by mutation. Thus, in this particular case, the capacity to transition to herbaceous may still exist in these plants, unless, of course, this rate class represents instances where clades have lost the genetic machinery to shut down cambial activity. Finally, in the medium ($q_{WM \rightarrow HM} = 0.0383$, s.e. $= \pm 0.0346$; $q_{WM \rightarrow HM} = 0.0012$, s.e. $= \pm 0.0149$) and fast ($q_{WF \rightarrow HF} = 99.8$, s.e. $= \pm 2.9$; $q_{HF \rightarrow WF} = 39.9$, s.e. $= \pm 1.9$) rate categories, it is far more likely for a lineage to transition from woody to herbaceous. The model also indicates that transitions among the different rate classes follow a general trend in which higher rates are inferred for transitions toward a slower rate class.

The general picture emerging within campanulids, particularly when "painting" the likeliest state and rate combinations onto internal nodes clearly supports the view
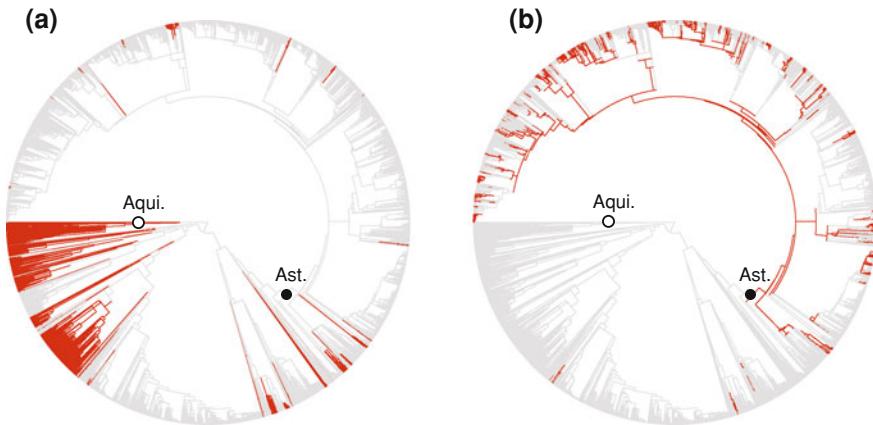
Fig. 16.3 Examples of branches "painted" as being in the different state and rate classes in the HRM with 3 rate classes applied to a phylogeny of 8,911 species of Campanulidae (campanulids) from Beaulieu et al. (2013a). Branches are *colored* based on whether the marginal probability is >0.75 (*dark*) of being in each growth habit state and rate class combination. The evolution of growth habit in campanulids clearly varies among clades. For example, woody clades such as Aquifoliales (denoted by an *open dot*) have slower, and the herbaceous clades such as Asteraceae (denoted by a *black dot*) generally have faster, rates of growth habit evolution, **a** woody, slow, **b** herb, medium + fast

that the rate at which growth habit evolves varies enough among clades to be biologically meaningful (Fig. 16.3). For example, most of the branches inferred to be in the faster rate classes appear to be confined only to Asteraceae, a geographically widespread clade (Bremer and Gustafsson 1997; Beaulieu et al. 2013b), where changes in growth habit as an adaptive response to new environments is well documented within the group (Carlquist 1974). In other groups, the evolution of growth habit is clearly limited, which may also point to their ecology, but also may reflect other additional underlying genetic factors. Within the Aquifoliales (e.g., hollies and their relatives), for example, the herbaceous habit has either never evolved or is a strategy that has not been successful, even though the group is currently widely distributed across both tropical and temperate regions. Whatever factors may be underlying these differences observed among clades, the use of the HRM clearly demonstrate that even though the capacity for growth habit evolution may be ever present, it is clearly expressed in fundamentally different ways.

16.6 Future Directions and Conclusions

Incorporating different rate classes of transition rates are not limited to binary characters. In a general sense, the concept of including "hidden" states can be used in any instance where the number of observed states is less than the number of

actual states. In other words, it is rather straightforward to extend the HRM to include characters that take on multiple states (i.e., >2 states). A good example comes from Maddison (1993): Consider a set of taxa whose observed states are red tails, blue tails, and no tails. In the HRM framework, one could treat this as having four hidden states—tail red, tail blue, tailless red, tailless blue—with three displayed states with tailless red and tailless blue both being present because tailless color is unknown. The HRM could then be used to address questions such as whether particular tailless species are more likely to have genes for red or blue color.

Similarly, biologists are not always interested in the transitions back and forth between states in just one binary character, but rather how the state of one binary character can affect the probability of change in another. There are ways to do this, of course, which have been highly influential (see Pagel 1994), but they still assume that the evolutionary process is homogeneous across the tree. Future extensions of the HRM will include models of correlated evolution between two or more binary characters in order to provide a new way of understanding the overall strength of character correlations and how it can change in various portions of phylogeny.

Future extensions will also include developing a HRM that, along with estimating transition rates, will also estimate the effect of a character state on speciation and extinction rates (i.e., BiSSE approach; Maddison et al. 2007; FitzJohn et al. 2009). Often these types of methods are used to test whether a character state is a "key innovation," as indicated by one state being correlated with higher net diversification rates (i.e., speciation–extinction) relative to another. At greater phylogenetic scales, however, the real effect between a character state and diversification rates is not always clear. What may seem like a causal connection may actually be due to other unmeasured factors or because the analysis included a nested clade that exhibits both the focal character and "something" else (Beaulieu and Donoghue 2013). The development of an HRM that includes the estimation of speciation and extinction rates as they relate to character states and rate classes would provide a powerful extension and allow for a more refined understanding of how particular character states influence the diversification process.

For now, the hidden rates framework can be used as a means of detecting differences in the evolution of a binary character and for the identification of models that can dramatically improve the fit to the underlying data. Unlike most existing approaches for both continuous and discrete characters, which require a priori assignment of models to different branches (O'Meara et al. 2006; Thomas et al. 2006; O'Meara 2007; Beaulieu et al. 2012), the HRM uses the observed character data directly to infer where the evolutionary model shifted in a phylogeny. In this way, the HRM is inherently exploratory, much like methods such as AUTEUR (Eastman et al. 2011) and SURFACE (Ingram and Mahler 2012; Mahler and Ingram Chap. 22 this volume) are for understanding clade-specific differences in continuous trait evolution. With the availability of such exploratory tools, we no longer have to restrict analyses, or describe results, in terms of a particular taxonomic rank (genus, family, order, etc.). Instead, we may begin to discover that

important evolutionary events better correspond to groups of taxa that do not have a formal name (e.g., Smith et al. 2011). It is in this way that methods such as the HRM will afford us with a far better understanding of evolution.

References

Abramowitz M, Stegun IA (1972) Handbook of mathematical functions, with formulas, graphs, and mathematical tables. National Bureau of Standards, Washington (DC)

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19:716–723

Beaulieu JM, Donoghue MJ (2013) Fruit evolution and diversification in campanulid angiosperms. Evolution 67:3132–3144

Beaulieu JM, Jhwueng D-C, Boettiger C, O'Meara BC (2012) Modeling stabilizing selection: expanding the Ornstein-Uhlenbeck model of adaptive evolution. Evolution 66:2369–2383

Beaulieu JM, O'Meara BC, Donoghue MJ (2013a) Identifying hidden rate changes in the evolution of a binary morphological character: the evolution of plant habit in campanulid angiosperms. Syst Biol 62:725–737

Beaulieu JM, Tank DC, Donoghue MJ (2013b) A Southern Hemisphere origin for campanulid angiosperms and traces of the break-up of Gondwana. BMC Evol Biol 13:80

Bell ET (1934) Exponential polynomials. Ann Math 35:258–277

Bremer K, Gustafsson MHG (1997) East Gondwana ancestry of the sunflower alliance of families. Proc Natl Acad Sci USA 94:9188–9190

Butler MA, King AA (2004) Phylogenetic comparative analysis: a modeling approach for adaptive evolution. Am Nat 164:683–695

Carlquist S (1974) Island biology. Columbia University Press, New York and London

Chase MW, Soltis DE, Olmstead RG, Morgan D, Les DH, Mishler BD et al (1993) Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbc*L. Ann Mo Bot Gard 80:528–580

Cronquist A (1968) The evolution and classification of flowering plants. Houghton Mifflin, Boston

Eastman JM, Alfaro ME, Joyce P, Hipp AL, Harmon LJ (2011) A novel comparative method for identifying shifts in the rate of character evolution on trees. Evolution 65:3578–3589

Felsenstein J (1981) A likelihood approach to character weighting and what it tells us about parsimony and compatibility. Biol J Linn Soc 16:183–196

Felsenstein J (2005) Using the quantitative genetic threshold model for inferences between and within species. Philos Trans R Soc B 360:1427–1434

Felsenstein J (2012) A comparative method for both discrete and continuous characters using the threshold method. Am Nat 179:145–156

Fitch WM, Markowitz E (1970) An improved method for determining codon variability and its application to the rate of fixation of mutations in evolution. Biochem Genet 4:579–593

FitzJohn RG, Maddison WP, Otto SP (2009) Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. Syst Biol 58:595–611

Galtier N (2001) A maximum-likelihood phylogenetic analysis under a covarion-like model. Mol Biol Evol 18:866–873

Groover AT (2005) What genes make a tree a tree? Trends Plant Sci 10:210–214

Holder MT, Lewis PO, Swofford DL (2010) The Akaike Information Criterion will not choose the no common mechanism model. Syst Biol 59:477–485

Huelsenbeck JP, Larget B, Alfaro ME (2004) Bayesian phylogenetic model selection using reversible jump Markov Chain Monte Carlo. Mol Biol Evol 21:1123–1133

Ingram T, Mahler DL (2012) SURFACE: detecting convergent evolution from comparative data by fitting Ornstein-Uhlenbeck models with stepwise Akaike Information Criterion. Methods Ecol Evol 4:416–425

Lens F, Smets E, Melzer S (2012) Stem anatomy supports *Arabidopsis thaliana* as a model for insular woodiness. New Phytol 193:12–17

Maddison WP (1993) Missing data versus missing characters in phylogenetic analysis. Syst Biol 42:576–581

Maddison WP, Midford PE, Otto SP (2007) Estimating a binary character's effect on speciation and extinction. Syst Biol 56:701–710

Marazzi B, Ane C, Simon MF, Delgado-Salinas A, Luckow M, Sanderson MJ (2012) Locating evolutionary precursors on a phylogenetic tree. Evolution 66:3918–3930

O'Meara BC, Ane C, Sanderson MJ, Wainwright PC (2006) Testing for different rates of continuous trait evolution. Evolution 60:922–933

O'Meara BC (2007) Estimating different rates of gene loss on a tree. Genetics 117:1415–1416

Pagel M (1994) Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proc R Soc B 255:37–45

Pagel M, Meade A (2004) A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Syst Biol 53:571–581

Pagel M, Meade A (2006) Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov Chain Monte Carlo. Am Nat 167(808):825

Penny D, McCormish BJ, Charleston MA, Hendy MD (2001) Mathematical elegance with biochemical realism: the covarion model of molecular evolution. J Mol Evol 53:711–723

Revell LJ (2008) On the analysis of evolutionary change along single branches in a phylogeny. Am Nat 172:140–147

Revell LJ (2014) Ancestral character estimation under the threshold model from quantitative genetics. Evolution 68:743–759

Smith SA, Beaulieu JM, Stamatakis A, Donoghue MJ (2011) Understanding angiosperm diversification using small and large phylogenetic trees. Am J Bot 98:1–12

Thomas GH, Freckleton RP, Szekely T (2006) Comparative analyses of the influence of developmental mode on phenotypic diversification rates in shorebirds. Proc R Soc Lond B 273:1619–1624

Tuffley C, Steel M (1997) Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bull Math Biol 59:581–607

Wright S (1934) An analysis of variability in the number of digits in an inbred strain of guinea pigs. Genetics 19:506–536

Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39:306–314

Yang Z, Kumar S, Nei M (1995) A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 141:1641–1650

Chapter 17
Detecting Phenotypic Selection by Approximate Bayesian Computation in Phylogenetic Comparative Methods

Nobuyuki Kutsukake and Hideki Innan

Abstract This chapter discusses the fundamental structure and advantages of the approximate Bayesian computation (ABC) algorithm in phylogenetic comparative methods (PCMs). ABC estimates unknown parameters as follows: (1) simulated data are generated under a suite of parameters randomly chosen from their prior distributions; (2) the simulated data are compared with empirical data; (3) parameters are accepted when the distance between the simulated and empirical data is small; and (4) by repeating steps (1)–(3), posterior distributions of parameters will be gained. Because ABC does not necessitate mathematical expression or analytic solution of a likelihood function, ABC is particularly useful when a maximum-likelihood (ML) estimation is difficult to conduct (a common situation when testing complex evolutionary models and/or models with many parameters in PCMs). As an application, we analysed trait evolution in which a specific species exhibits an extraordinary trait value relative to others. The ABC approach detected the occurrence of branch-specific directional selection and estimated ancestral states of internal nodes. As computational power increases, such likelihood-free approaches will become increasingly useful for PCMs, particularly for testing complex evolutionary models that deviate from the standard models based on the Brownian motion.

N. Kutsukake (✉) · H. Innan
Department of Evolutionary Studies of Biosystems, The Graduate University
for Advanced Studies, Kanagawa Hayama 240-0193, Japan
e-mail: kutsu@soken.ac.jp

N. Kutsukake · H. Innan
PRESTO, Japan Science and Technology Agency, 4-1-8 Honcho Kawaguchi,
Saitama 332-0012, Japan

17.1 Background

> Evolution is messy. Rates, direction, and mode of evolution vary through time and among
> clades and characters, and this inconstancy itself will often be unpredictable and haphazard.
> Phylogenetic methods that ignore this variation will often produce inaccurate and misleading
> results. As a result, researchers must embrace statistical approaches that assess such variation
> rather than assuming constancy (Losos 2011).

Phylogenetic comparative methods (PCMs) are powerful approaches to test evolutionary models of speciation and trait evolution (Chap. 1). Although PCMs have been widely used in evolutionary biology, it is always important to remember that statistical inferences regarding evolutionary parameters are based on assumptions and hypotheses. In studies of continuous traits, single-rate Brownian motion (BM) has been commonly used as a model for character evolution. BM represents a neutral evolution or evolution tracking a continuously fluctuating optimum value and is a good approximation for describing a *pattern* of trait divergence. However, this does not mean that the BM-like evolutionary mode can always approximate the *process* of divergence well (Estes and Arnold 2007; Gingerich 2009). It is natural to assume that the process comprises heterogeneous evolutionary rates and modes; that is, they vary within lineages, among branches, and within clades (Losos 2011). One approach to coping with this heterogeneity is to test a multiple-rate BM model against a simple single-rate BM model. Researchers can set a model with varying rates for each branch or for monophyletic subgroups based on a priori biological hypothesis (O'Meara et al. 2006; Thomas et al. 2006), or set a model without specifying a hypothesis (Venditti et al. 2011). Another approach includes scaling parameters on branch lengths or evolutionary rates, and test models of punctuated, accelerating/decelerating, or early burst evolution (Table 17.1; Blomberg et al. 2003; Pagel 1997, 1999; Harmon et al. 2010). Additionally, the local occurrence of stabilising selection can be modelled by using the Ornstein–Uhlenbeck (OU) process (Table 17.1; Chap. 15). Parameters specific to each model (Table 17.1) are incorporated to calculate an expected covariance among the traits of each species. The variance–covariance matrix, a critical component of the likelihood function of PGLS (phylogenetic generalised least squares), is used for estimating unknown parameters via a maximum-likelihood (ML) estimation or a Bayesian approach. With these approaches, it is currently possible to address a wide range of questions. However, some methods are mathematically complex and not always transparent for general users of PCMs. Moreover, traditional approaches may not be well enough established to test complex evolutionary scenarios with many parameters because the description of the variance–covariance matrix is not straightforward to gain. It is preferable to have a flexible toolkit that allows for testing of such complex evolutionary models.

A simulation-based likelihood approach using an approximate Bayesian computation (ABC) (Bokma 2010; Slater et al. 2012; Kutsukake and Innan 2013) may

Table 17.1 Examples of parameters used in previous studies of PCMs

Model	Parameters	Biological meaning
Scaling branch length	λ[a]	Phylogenetic signal
	δ[a]	Temporal rate change
	κ[a]	Gradual versus punctuational evolution
ACDC	g[b]	The overall rate of acceleration (AC) or deceleration (DC)
Early burst	r[c]	The pattern of rate change through time
OU (whole phylogeny)	d[b]	A restraint parameter in the OU transformation
OU	α[d]	Strength of stabilising selection

[a] Pagel (1997, 1999)
[b] Blomberg et al. (2003)
[c] Harmon et al. (2010)
[d] Hansen (1997), Butler and King (2004)

be one solution. This approach is useful when analytic expressions of likelihood and the ML estimator cannot be gained. In this chapter, we aim (1) to explain the basic structure, method, and caveats of ABC, (2) review the application of ABC to PCMs, (3) provide a protocol for the ABC approach, (4) provide an example using ABC, and (5) discuss the future direction of the ABC approach.

17.2 Approximate Bayesian Computation

The ABC framework was originally developed in population genetics (Tavare et al. 1997) and gradually introduced to the disciplines of ecology and evolutionary biology (Beaumont 2010; Bertorelle et al. 2010; Csillery et al. 2010). The major advantage of ABC is that parameter estimation can be performed even when the likelihood of the data cannot be computed, most often due to data complexity. The fundamental structure of ABC is as follows: let x be the observed data and assume an evolutionary model with the parameters η, which we aim to estimate. η could be a vector with multiple parameters.

(1) Determine the prior distributions of all parameters η in x.
(2) For each parameter, generate a random value from the prior distribution. A random set of the parameters at the ith simulation is denoted by η_i'.
(3) Simulate data x_i' using η_i'.
(4) Accept η_i' if x_i' is identical to the observed data x.
(5) Go to (2) until a large number of accepted η' values have been accumulated.

In most cases, the simulation will rarely produce x_i' that is completely identical to the observed data x. To solve this problem, instead of using the full data set, ABC usually employs summary statistics (denoted by s, which usually comprises several types of summary statistics, i.e., $s = [S_1, S_2, \ldots S_n]$) and use $s(x)$ and $s(x_i')$ instead of x and x_i'. Even when summary statistics are used, it may be rare to gain

an $s(x_i')$ value that is exactly the same as $s(x)$. In such a case, we can set a certain range of tolerance. One simple algorithm, a rejection sampling, uses the following relaxed criterion:

$$(4') \text{ Accept } \eta_i' \text{ when } \rho(s(x) - s(x_i')) < \varepsilon,$$

where ε is a tolerance and $\rho(\cdot)$ is a function for calculating a distance between $s(x)$ and $s(x_i')$, representing their similarity. By setting a tolerance, the efficiency of parameter acceptance will dramatically increase in comparison with (4). In addition to this simple rejection sampling algorithm, more sophisticated methods have been proposed to improve data-acceptance efficiency, such as importance sampling or ABC–MCMC (Markov chain Monte Carlo; see Marjoram et al. 2003 and Majoram and Tavare 2006 for details).

17.3 Caveats of ABC

Although Sect. 17.2 provided the fundamental structure of ABCs, several general caveats remain that users should be aware of. Because of space limitations, we will briefly discuss three core points, summary statistic sufficiency, estimation robustness, and model selection, although other general caveats of Bayesian computation apply (see Gelman et al. 2013).

First, the number and choice of summary statistics is critical in ABC (see Beaumont et al. 2002; Csillery et al. 2010; Leuenberger and Wegmann 2010 for details). The use of summary statistics reduces the dimensionality and complexity of the data (Tavare et al. 1997). Accordingly, if researchers use too few summary statistics, too much information is discarded, resulting in a low resolution of the parameter estimation. However, if too many summary statistics are used, the acceptance rate will be so low that it will be computationally intensive. Researchers must examine the performance of their chosen summary statistics before applying ABC to real data, such as by conducting power tests with artificially generated data.

The second point concerns estimation robustness and computational efficiency determined by the level of tolerance ε. If tolerance is not sufficiently small, parameter estimation will be rough (Beaumont 2010), while too severe tolerance is not practical because computational efficiency will be too low. Therefore, researchers are required to set an adequate level of tolerance in ABC, meaning that a certain amount of subjectivity and uncertainty remains in parameter estimation. To address this problem, post hoc correction approaches have been proposed to increase the accuracy of parameter estimation, such as post-sampling regression adjustment using general linear model (ABC–GLM; Leuenberger and Wegmann 2010).

Third, simple model selection cannot be used in the framework of ABC. This difficulty stems from the fact that likelihoods are not calculated, and data acceptance in each model depends on summary statistics and tolerance (Beaumont 2010;

Csillery et al. 2010; Robert et al. 2011). Although this can be problematic in the case of comparing un-nested models, model selection based on posterior distributions is relatively straightforward when models are nested (explained below).

17.4 Application of ABC to PCMs

Three PCM studies have used ABC for analysing trait evolution (Table 17.2; see Rabosky (2009) for an analysis of clade diversification using ABC). These studies aimed to test evolutionary models that are difficult or not straightforward to handle in the traditional framework of PCMs.

The first PCM study using ABC was performed by Bokma (2010). This study aimed to separate the effects of parameters of cladogenetic evolution and of anagenetic evolution on trait disparity. There are wide variations in trait disparity among clades, and understanding the ecological and evolutionary causes of this phenomenon has been a central interest in macroevolutionary studies. It is likely that a trait disparity is positively correlated with a parameter of anagenetic evolution (i.e. rate parameter of BM). Additionally, the number of speciation events may be positively correlated with the trait disparity because the interval between speciation events should be smaller in larger clades, automatically resulting in larger trait variance (Ricklefs 2004, 2006; Purvis 2004). Thus, the effects of anagenetic and cladogenetic evolution on trait disparity are confounded and difficult to separate. To solve this problem, Bokma (2010) applied ABC to trait disparity in passerine data (originally used in Ricklefs 2004) to investigate the relative importance of those parameters with respect to anagenetic evolution, cladogenetic evolution, and their combination. Bokma (2010) conducted a simulation using a model comprising two free parameters (anagenetic and cladogenetic parameters) and estimated those parameters using phenotypic variance as summary statistics. They found that gradual anagenetic change was more important than a combination of anagenetic and cladogenetic evolution in terms of explaining the phenotypic divergence observed in this group.

In another study, Slater et al. (2012) used a mixture of Markov-chain Monte Carlo (MCMC) and ABC to infer the speciation/extinction rates and parameters of trait evolution in an incompletely sampled phylogeny. This study was motivated by a common problem in PCM studies: it is usually difficult to perform complete sampling of a phylogeny and collect trait data from all species. Data incompleteness could reduce the accuracy of statistical inference of evolutionary parameters and the power to test models. Nevertheless, ABC can handle such incomplete data relatively well because data are transformed into summary statistics so that it is not always necessary to sample all data (as long as collected data are unbiased and represent each clade). Using this unique characteristic of ABC, Slater et al. (2012) developed a new framework with which to test evolutionary models using incomplete data. First, speciation and extinction parameters were

Table 17.2 Summary of three PCM studies using ABC

Study	Parameters	Reason for using ABC	Summary statistics	Empirical example
Bokma (2010)	Anagenetic and cladogenetic evolution parameters	No analytical solution for an evolutionary model	Phenotypic variance	Parameter estimation of phenotypic variance in passerine birds
Slater et al. (2012)	Speciation rate, extinction rate, ancestral state, BM rate	Incompletely sampled phylogeny and trait data	Mean and variance for terminal lineages	Comparison of terrestrial carnivores and pinnipeds
Kutsukake and Innan (2013)	See Table 17.3 for details	Branch-specific directional selection	Direct (and full) likelihood	Detection of directional selection in human brain size

estimated from known data by MCMC. They conducted simulations to generate a phylogeny under a birth/death process using sampled parameter values. Next, using ABC, they simulated trait evolution according to BM rates on the generated phylogeny using the rate parameter(s) of BM and sampled the trait value at the root. Finally, summary statistics (mean and variance) calculated from the simulated trait data were compared with those calculated from the real data of each clade. This study applied this algorithm to test the difference in BM rate parameters of the body size between pinnipeds and terrestrial carnivores. This study was interested in estimating two to three parameters, namely the trait value at the root, and one or two rate parameters assigned to those two groups. Given that a model with one rate parameter is nested within a model with two rate parameters, this study used a posterior probability to select the two models. The ABC approach revealed that, contrary to expectation, rate parameters did not differ between the two groups.

Testing multi-BM models is also possible in the traditional framework of PCMs (O'Meara et al. 2006; Thomas et al. 2006), but it requires the assumption of complete sampling. Slater et al. (2012) demonstrated that ABC handles this technical problem quite well and stated that their method is applicable to other existing evolutionary models, thereby providing an important first step in testing complex evolutionary models by ABC. This method was implemented in the software *MECCA* (Modeling Evolution of Continuous Characters using ABC)

Third, our recent study (Kutsukake and Innan 2013) used ABC under an evolutionary model with heterogeneous evolutionary modes and rates within a phylogeny (assuming that the phylogeny is known). Although this algorithm is designed to model wide ranges of trait evolution, one useful application is to detect directional selection occurring locally within the phylogeny. In molecular evolution studies, there is a popular approach to measure branch-specific directional selection (e.g. dN/dS or Ka/Ks) (Li 1997; Yang 2006). Motivated by this approach, our model was designed to incorporate branch-specific selection parameters and

allowed statistical testing and measurement of the intensity of selection. We considered the BM model (with slight modifications) to be a null neutral model. This simplest model can be extended by adding as many branch-specific selection parameters as desired for setting an alternative.

Using this framework, Kutsukake and Innan (2013) showed a simple example analysis of data on brain volume among four species of great apes. In total, three parameters were involved: the trait value at the most recent common ancestor, the background evolutionary rate, and the strength of directional selection at the human lineage (the method will be explained later). It was shown that the trait evolution on the branch reaching to humans significantly deviated from the BM mode, exhibiting strong evidence of branch-specific directional selection.

Although only a few PCM studies have used ABC thus far, ABC has a great potential to be applied to a wide range of problems and settings. Since phenotypic evolution consists of a process in which trait values increase or decrease, listing all parameters making up this process should be sufficient to build any complex evolutionary models. Of PCM studies using ABC, only our model is flexible enough to be applied to various evolutionary processes at any time point; these can be deviate from the simple BM process and include various modes and intensities of selection on different branches. Another strength is that our framework employs direct likelihood as a summary statistic, by which it is possible to avoid the problem of which and how many summary statistics should be used in ABC (see Sect. 17.3). Furthermore, intraspecific variation, a factor that has been overlooked but is known to affect to parameter estimation in PCMs (Garamszegi and Møller 2010; Chap. 7), and uncertainty in phylogeny can be taken into account (see below and Table 17.3). Below, we describe the model and algorithm of our approach in detail.

17.5 ABC Algorithm in Kutsukake and Innan (2013)

We herein explain the algorithm of Kutsukake and Innan (2013) in more details (see also Fig. 17.1a for a simplified structure of this algorithm; an example program written in the C language in shown in the Online Practical Material, http://www.mpcm-evolution.com).

A necessary data set for applying this algorithm is the same as those for other studies. First, trait data are required for each species. Intraspecific variation can also be incorporated by setting a distribution of the trait. Any kind of distribution can be handled, from a regular quantitative trait that likely follows a simple normal distribution to a trait with a discrete distribution. The phylogeny ψ (topology and branch length τ) of the species is also needed and is assumed to be known (this assumption could be relaxed as phylogenetic uncertainty can be taken into account in the algorithm; see Table 17.3). Λ represents all other parameters involved. At minimum, may comprise the trait value of the most recent common ancestor (θ_0) and evolutionary rate μ. The evolutionary rate, the number of evolutionary events (i.e. mutation and

Table 17.3 Main parameters used in a simulation-based likelihood approach by Kutsukake and Innan (2013)

Parameter	Biological meaning	Notes
Known parameters		
ψ	Species tree (topology and branch lengths)	When topology and/or branch length includes uncertainty, it is possible to consider those uncertainties by using a randomly chosen topology and/or a set of random values for branch length in each simulation
τ_i	Length of the ith branch on the tree in a given unit (e.g. time or genetic distance)	
n	Number of nodes	
$\Omega = \{\Omega_1, \Omega_2, \Omega_3, \ldots \Omega_n\}$	Observed nodal data (e.g. mean)	This setting assumes data contain no measurement errors. Measurement errors can be incorporated by considering that σ contains both intraspecific variation and measurement errors, or by setting a new parameter set to represent measurement error. Internal nodes such as fossil data can also be used
$\sigma = \{\sigma_1, \sigma_2, \sigma_3, \ldots \sigma_n\}$	Observed nodal data on intraspecific variation (standard deviation or other parameters)	
A parameter set of interests (Λ)		
θ_0	Phenotype of the MRCA	
μ	Evolutionary rate (the number of a phenotypic change) per time unit (e.g. a million years, a generation) and genetic distances	μ^+ and μ^- for increasing or decreasing a trait value. It is possible to assume different evolutionary rates for each branch i by setting μ_i^+, μ_i^-
Other factors		
ϕ	Changes in phenotypic values caused by a single event of evolution; this parameter can be set arbitrarily	ϕ^+, ϕ^- for different effects for increasing or decreasing trait value. For example, it is possible to model a larger evolutionary effect increasing a trait than one decreasing a trait by setting $\phi^+ > \phi^-$. In the case that the evolutionary effect differs among branches, a different value (ϕ_i^+, ϕ_i^-) can be used

(continued)

Table 17.3 (continued)
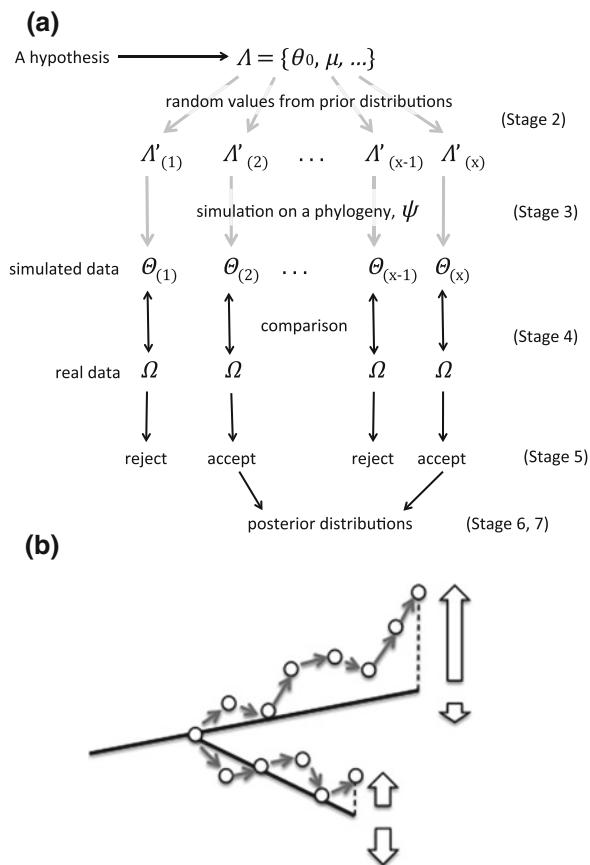
Parameter	Biological meaning	Notes
$s_i = f(\mu, \tau)$ or $f(\mu, \tau, \ldots)$	Number of fixed mutations that increase (s_i^+) or decrease (s_i^-) the phenotypic value at a certain branch i	A random integer from a Poisson distribution with mean $\mu_i^+ \tau_i$ (or $\mu_i^- \tau_i$) in the case of BM evolution. If one wants to test more complex evolutionary rate in which the evolutionary rate is a function of the branch length or time from root, such as accelerating or decelerating evolutionary rate, additional parameters can be used in this function
$\Theta = \{\theta_1, \theta_2, \theta_3, \ldots \theta_n\}$	Simulated data of n descendent species	

fixation), is denoted by μ_i at branch i, whose length is τ_i. Thus, the expected number of evolutionary events increasing and/or decreasing the trait value is $\mu_i \tau_i$. The effect of each evolutionary event on the trait is denoted by ϕ, which should follow a certain distribution. By setting branch-specific $\mu_i \tau_i$, it is possible to model the density distribution of the phenotypic change for each branch and to test wide ranges of evolutionary models such as branch- or clade-specific directional selection (see below) or the OU process (see Kutsukake and Innan 2013 for details). This setting enables researchers to treat those models and a null model (neutral evolution) as nested models, allowing one to avoid the complicated problem of model selection in ABCs (see Sect. 17.3).

Under this framework, it is possible to apply the ABC algorithm along the following four steps. The steps correspond to the general ABC procedure in Sect. 17.2.

Step 1 *Determine the prior distribution of each parameter.* An advantage of Bayesian statistics is that it enables the setting of informative (i.e. strong) prior distributions based on prior biological knowledge.

Step 2 *Choose a random value for each parameter from the prior distribution.* Parameters used in the simulation (Λ') are randomly chosen from their prior distributions.

Step 3 *Let trait values evolve by simulation.* Simulation of the trait evolution of phylogeny ψ is conducted using Λ'. As a result, simulated values Θ of traits of n species are gained.

Step 4 *Calculate the likelihood by comparing simulated data with the real data and determine whether that parameter set is accepted or rejected.* Each simulated value θ_i is compared with the real value Ω_i. By comparing n species, a joint probability (full likelihood) $\Pr(\Omega|\Theta) = \Pr(\Omega_1, \Omega_2, \Omega_3, \ldots, \Omega_n|\theta_1, \theta_2, \theta_3, \ldots, \theta_n)$ will be calculated. In the case that this probability is computationally very difficult to gain, a composite

Fig. 17.1 a An illustrative example of the ABC approach. A trait simulation can be conducted based on a parameter set (Λ) on a phylogeny with a discrete timescale. The number of simulation rounds is shown in parenthesis. A number (x) of simulations were conducted until enough samples were collected to infer posterior distribution. **b** The trait value can be simulated by the BM-like evolutionary mode (a *lower branch*) or directional selection to increase trait value (an *upper branch*). Vertical axis indicates a trait value, and *white arrows* indicate direction of trait evolution for each branch, with its length corresponding to the number of evolutionary events increasing or decreasing trait values

(a)

A hypothesis $\longrightarrow$ $\Lambda = \{\theta_0, \mu, ...\}$

random values from prior distributions

(Stage 2)

$\Lambda'_{(1)}$ $\quad$ $\Lambda'_{(2)}$ $\quad$ ... $\quad$ $\Lambda'_{(x-1)}$ $\Lambda'_{(x)}$

simulation on a phylogeny, ψ $\qquad$ (Stage 3)

simulated data $\quad$ $\Theta_{(1)}$ $\qquad$ $\Theta_{(2)}$ $\quad$ ... $\qquad$ $\Theta_{(x-1)}$ $\Theta_{(x)}$

comparison $\qquad$ (Stage 4)

real data $\quad$ Ω $\qquad$ Ω $\qquad\qquad$ Ω $\qquad$ Ω

reject $\quad$ accept $\qquad\qquad$ reject $\quad$ accept $\qquad$ (Stage 5)

posterior distributions $\qquad$ (Stage 6, 7)

(b)

likelihood $\Pr(\Omega|\Theta) = \prod_{i=1}^{n} \Pr(\Omega_i|\theta_i)$ may be used as an approximate proxy, which should work fairly well as long as the sampled species are reasonably diverged. Here, one important advantage in our framework is that the choice of summary statistics is avoided. Although previous studies have used means or variance as summary statistics (Table 17.2), this study employed a more straightforward method by using the direct likelihood instead of a summary statistic. The use of likelihood as a summary statistics may not be common in standard ABC, where data are usually represented by a set of summary statistics and likelihood cannot be analytically computed. Note that we can compute the likelihood of the observed data given a "simulated data set". In our framework, given a parameter set, a single run of random simulation provides a simulated data set representing a single realisation of the random process. Then, the likelihood is computed given this simulated data set. This is different from the standard ML estimation that requires the likelihood given a parameter

set. Intraspecific variation in the trait data can be considered by using the probability density function when calculating $\Pr(\Omega_i|\theta_i)$, the probability of gaining Ω given θ. Thus, by assuming a certain distribution of intraspecific variation, we can evaluate the similarity between the simulated and real data using the likelihood. As a consequence, our algorithm can avoid the common problem of choosing appropriate summary statistics.

Using $\Pr(\Omega|\Theta)$, acceptance of Λ' can be determined. As described above, there are several methods of judgment (Marjoram et al. 2003; Majoram and Tavare 2006); the researchers determine which method will be used. However, we caution users on choosing the acceptance threshold. A strict threshold increases the precision of parameter estimation, but if it is too strict, the computational load may be too large. A reasonable choice is desired so as to gain posterior distributions within a realistic computation time (see Sect. 17.3).

Step 5 *Obtain posterior distributions of the parameters and assess the importance of each parameter.* Repeat Steps (2)–(4) until a sufficient number of parameters is accepted. This usually requires intensive iterative computation, particularly when there are many parameters. Posterior distributions and credible intervals can be used to judge whether each parameter is different from a specific value (e.g. a value expected under a null model) and to accordingly test which evolutionary models are supported.

17.6 Detecting Branch-Specific Directional Selection

In this ABC approach, it is possible to let phenotypic traits evolve via directional selection only in certain branches (Fig. 17.1b). There are various ways to model such local occurrence of directional selection. Kutsukake and Innan (2013) introduced a simple model in which the direction and intensity of selection is parameterised by a single parameter, k. That is, it is assumed that if selection favours the increase in a trait at a branch with $\tau\mu$, the expected numbers to increase and decrease the trait are given by $k\tau\mu$ and $\tau\mu/k$, respectively. When k equals 1, the result is the same as neutral evolution, as the numbers of evolutionary events increasing and decreasing the trait value are identical on average.

The model using k is not the only universal way to model directional selection. Other types of directional selection can be flexibly modelled by setting new parameters. For example, directional selection in one direction can co-occur with constant occurrence of neutral evolution, including evolution in the opposite direction. In such cases, researchers can set a new parameter, and evolutionary events increasing a trait value can be multiplied by that parameter while those decreasing the trait value will be untouched. In other cases, consistent directional selection of a trait may start to operate from a certain intermediate point of a branch because of alterations in a fitness landscape by environmental changes. In such cases, researchers can create a model in which a trait value evolves neutrally

until the intermediate point of the branch, and the evolutionary rate is multiplied by a parameter of directional selection in the rest of the branch.

Given that there are several ways to model directional selection and to represent the relative intensity of selection, it is useful to quantify the expected change in the trait value for each branch, which is denoted by Δ_{sel}. In the case of using k, the Δ_{sel} value can be calculated by $\mu_i^+ \tau_i k - \mu_i^- \tau_i / k$ (Kutsukake and Innan 2013). To test the presence of directional selection, the posterior distribution of Δ_{sel} can be compared to zero or to a maximum value of the trait change under a pure neutral model (denoted by Δ_{null} in Kutsukake and Innan 2013). The advantage of using Δ_{sel} is that it can cancels out the confounding relationship between μ and k (or other parameters for directional selection); that is, k is inversely correlated with μ because those two parameters compensate for each other to let a phenotypic value increase (or decrease) towards a certain value. In other words, a low evolutionary rate must necessitate strong directional selection, whereas weak directional selection may be sufficient when evolutionary rate is large. In such a situation, the most meaningful quantity should be Δ_{sel} rather than μ or k.

17.7 Application to Weevil Rostrum Demonstrating Branch-Specific Direction Selection

We analysed weevil rostrum evolution to show how our framework can be applied to a complex model of trait evolution with branch-specific directional selection (see OPM). Toju and Sota (2006) studied interspecific variation in the rostrum, an organ used to excavate host plant fruits to lay eggs inside, among seven species of weevils. They investigated two nearly isolated subpopulations of one of those species, *Curculio camelliae*. Interestingly, their interpopulation comparison between the subpopulations showed a positive correlation between the thickness of the camellia fruit walls and the weevil rostrum length, suggesting that an arms race occurred between those two traits. Toju and Sota (2006) applied PGLS (Chaps. 5 and 6) and showed significant effects of two scaling parameters, κ and δ (Pagel 1999). The significant effects of these parameters indicate punctuated evolution and the large effect of a short branch. They also estimated the ancestral states by the ML approach (Schulter et al. 1997). As a result, the rostrum length of the common ancestor (an internal node C in Fig. 17.2a) between *C. camelliae* and its sister species (*C.* species on *C. sasanqua*) was estimated at 8.12, an approximately intermediate value of their descendent species (Fig. 17.2a). This estimated value suggests that the rostrum length should have increased in *C. camelliae* but decreased in its sister species.

Note that these results reflect BM-based models, in which a uniform rate of evolution was applied to the entire tree, and directional selection on the lineage of *C. camelliae* was not specifically incorporated. Therefore, the estimated ancestral states of *C. camelliae* may have been overestimated. We re-analysed the data

reported by Toju and Sota (2006) and tested the model showing that branch-specific directional evolution has increased the rostrum length in the lineage leading to *C. camelliae*. In the ABC, we conducted trait simulation by setting three parameters: the ancestral value, background (neutral) evolutionary rate, and intensity of directional selection k. We assumed a Gaussian distribution of rostrum length and calculated the probability for the ith species by considering intraspecific variation σ as follows:

$$\Pr(\Omega_i|\theta_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left[-\frac{(\theta_i - \Omega_i)^2}{2\sigma_i^2}\right]. \qquad (17.1)$$

We accepted data by the criterion that the probability of acceptance is proportional to a direct likelihood (Kutsukake and Innan 2013). Using this equation, we can avoid the common problem of the tolerance (see Sect. 17.3).

The obtained posterior distributions favoured our model of branch-specific directional selection over the null model based on neutral evolution. The posterior distribution of k did not overlap with 1, the value indicating neutral evolution (Fig. 17.2b). The ancestral value estimated by this branch-specific directional selection model was much smaller than that estimated by ML estimator, not only in the ancestor of *C. camelliae* but also in other internal nodal species (Fig. 17.2b). These differences are obviously due to the incorporation of branch-specific directional selection. This estimation supports the idea that a co-evolutionary arms race promoted the evolution of an exaggerated trait.

Although an OU model can also be applied to the branch to the Japanese species, we believe that applying directional selection is more suitable in this case because there should be no adaptive optimum in an arms race, and phenotypic shifts should be unidirectional.

17.8   Further Applications

An advantage of the ABC approach is that researchers can flexibly test complicated evolutionary models even when their likelihood cannot be computed. This advantage meets recent demands of PCMs as comparative approaches are currently applied to wide ranges of biological problems and complicated models. Another advantage of ABC is that trait simulation provides a good opportunity to critically consider each stage of an evolutionary event. Using simulations, researchers can create models focusing on evolutionary processes, rather than evolutionary patterns. However, user should be aware of the caveats of ABCs (discussed in Sect. 17.3). Careful calibration of summary statistics, tolerance, power analysis, and well-designed model settings are necessary. The ABC approach itself is also rapidly developing, and we recommend that ABC users follow the ongoing improvements and debates.
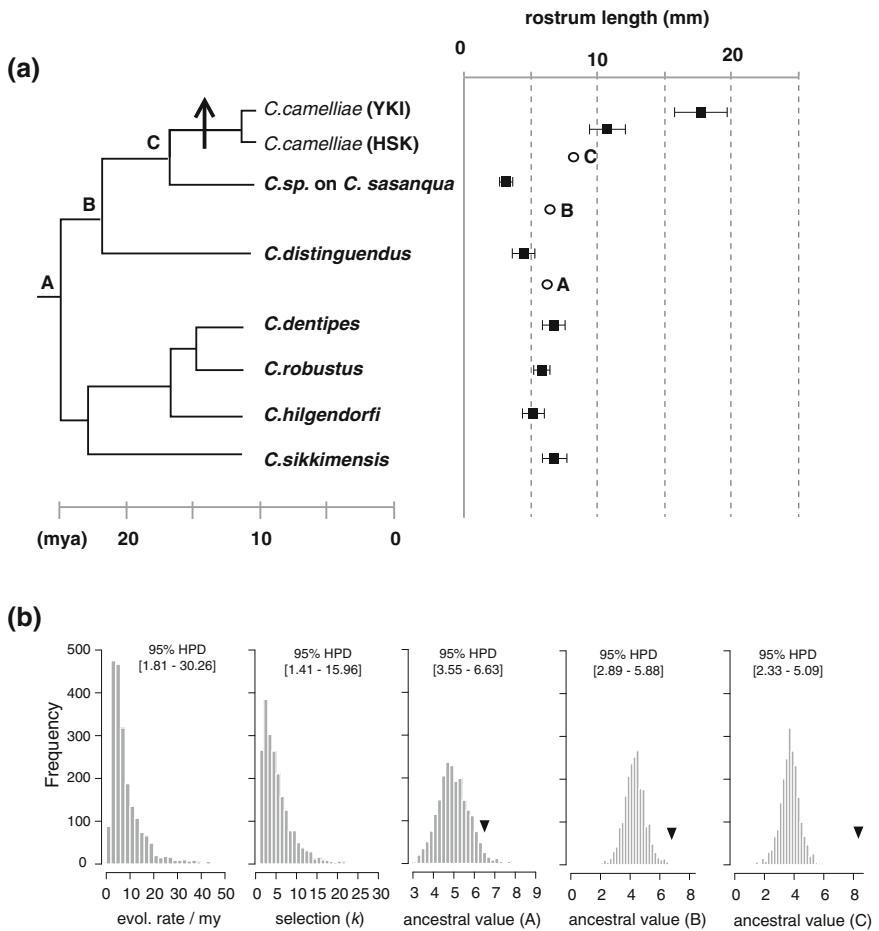
Fig. 17.2 a Phylogeny and trait values (rostrum length; mean and 1 SE) of seven species of weevils (*C. camelliae* were sampled in two locations: YKI and HSK). Phylogenetic relationship of these species was estimated by mitochondrial COI gene sequences (Toju and Sota 2006). The trait values indicated by *white circles* correspond to the estimated values of internal nodes *A*, *B*, and *C* by Toju and Sota (2006). We tested an evolutionary model in which directional selection to increase a trait value has occurred between internal nodes *C* of the ancestor of *C. camelliae* (indicated by a *thick line* with an *upward arrow*). Simulations were performed by a discrete timescale using a million years. **b** Posterior distributions of parameters and 95 % confidence intervals based on 2,000 accepted parameter sets (ancestral value at internal nodes *A*, *B*, and *C*, evolutionary rate, and selection) of our ABC analysis. We did not accept simulated data with a negative trait value. *Black inverted triangles* indicate trait values estimated by Toju and Sota (2006). In this analysis, ϕ was set as an exponential distribution with a mean of 0.05 mm, meaning that one evolutionary change results in a trait change whose magnitude is a random value from an exponential distribution whose average is 0.05 mm. Prior distributions were set as follows: MRCA $\sim U(3.15, 9)$, evolutionary rate per a million year $\sim U(0, 50)$, and $k \sim U(0.0001, 30)$, where $U(\cdot)$ indicates a uniform distribution

We believe that ABC-based PCMs have many developmental possibilities and can be applied to broad ranges of evolutionary questions and empirical data. As our example showed, one can model branch-specific directional selection by setting a specific value of the evolutionary rate (μ; Table 17.3) at each branch. As an extension of this directional selection model, an acceleration of selection pressure, often witnessed during co-evolutionary arms races, can be also analysed. With further modifications ϕ (Table 17.3), one can model an evolutionary scenario in which different species have a different degree of trait change stemming from one evolutionary event. Such a situation is common in analyses of size traits or body mass, in which the degree of trait change (e.g. an increase/decrease in body mass) positively correlates with its species trait value (e.g. body mass). Furthermore, changes in the evolutionary rate and mode in the middle of a branch, hybridisation, and cultural evolution can also be flexibly incorporated in this framework.

Finally, we should note that our ABC approach may not fit to a ready-made software or statistical library because flexibility is the most important advantage of the ABC. It is ideal that each user writes programs for evolutionary models that the given researcher would like to test (see OPM).

References

Beaumont MA (2010) Approximate Bayesian computation in evolution and ecology. Ann Rev Ecol Evol Syst 41:379–406

Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. Genetics 162:2025–2035

Bertorelle G, Benazzo A, Mona S (2010) ABC as a flexible framework to estimate demography over space and time: some cons, many pros. Mol Ecol 19:2609–2625

Blomberg SP, Garland T Jr, Ives AR (2003) Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution 57:717–745

Bokma F (2010) Time, species, and separating their effects on trait variance in clades. Syst Biol 59:602–607

Butler MA, King AA (2004) Phylogenetic comparative analysis: a modeling approach for adaptive evolution. Am Nat 164:683–695

Csillery K, Blum MG, Gaggiotti OE, François O (2010) Approximate Bayesian computation (ABC) in practice. Trends Ecol Evol 25:410–418

Estes S, Arnold SJ (2007) Resolving the paradox of stasis: models with stabilizing selection explain evolutionary divergence on all timescales. Am Nat 169:227–244

Garamszegi LZ, Møller AP (2010) Effects of sample size and intraspecific variation in phylogenetic comparative studies: a meta-analytic review. Biol Rev 85:797–805

Gelman A, Carlin JB, Stern HS, Rubin DB (2013) Bayesian data analysis, 3rd edn. CRC Press, London

Gingerich PD (2009) Rate of evolution. Ann Rev Ecol Evol Syst 40:657–675

Hansen TF (1997) Stabilizing selection and the comparative analysis of adaptation. Evolution 51:1341–1351

Harmon LJ, Losos JB, Jonathan Davies T, Gillespie RG, Gittleman JL et al (2010) Early bursts of body size and shape evolution are rare in comparative data. Evolution 64:2385–2396

Kutsukake N, Innan H (2013) Simulation-based likelihood approach for evolutionary models of phenotypic traits on phylogeny. Evolution 67:355–367

Leuenberger C, Wegmann D (2010) Bayesian computation and model selection without likelihoods. Genetics 184:243–252

Li W-H (1997) Molecular evolution. Sinauer Associates, Sunderland, Massachusetts

Losos JB (2011) Seeing the forest for the trees: the limitations of phylogenies in comparative biology. Am Nat 177:709–727

Marjoram P, Tavare S (2006) Modern computational approaches for analysing molecular genetic variation data. Nat Genet Rev 7:759–770

Marjoram P, Molitor J, Plagnol V, Tavare S (2003) Markov chain Monte Carlo without likelihoods. Proc Natl Acad Sci USA 100:15324–15328

O'Meara BC, Ane CM, Sanderson MJ, Wainwright PC (2006) Testing for different rates of continuous trait evolution in different groups using likelihood. Evolution 60:922–933

Pagel M (1997) Inferring evolutionary processes from phylogenies. Zool Scr 26:331–348

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

Purvis A (2004) Evolution: how do characters evolve? Nature 432:166

Rabosky DL (2009) Heritability of extinction rates links diversification patterns in molecular phylogenies and fossils. Syst Biol 58:629–640

Ricklefs RE (2004) Cladogenesis and morphological diversification in passerine birds. Nature 430:338–341

Ricklefs RE (2006) Time, species, and the generation of trait variance in clades. Syst Biol 55:151–159

Robert CP, Cornuet JM, Marin JM, Pillai NS (2011) Lack of confidence in approximate Bayesian computation model choice. Proc Natl Acad Sci USA 108:15112–15117

Schulter D, Price T, Mooers AO, Ludwig D (1997) Likelihood of ancestor states in adaptive radiation. Evolution 51:1699–1711

Slater GJ, Harmon LJ, Wegmann D, Joyce P, Revell LJ, Alfaro ME (2012) Fitting models of continuous trait evolution to incompletely sampled comparative data using approximate Bayesian computation. Evolution 66:752–762

Tavare S, Balding DJ, Griffiths RC, Donnelly P (1997) Inferring coalescence times from DNA sequence data. Genetics 145:505–518

Toju H, Sota T (2006) Phylogeography and the geographic cline in the armament of a seed-predatory weevil: effects of historical events vs. natural selection from the host plant. Mol Ecol 15:4161–4173

Thomas GH, Freckleton RP, Szekely T (2006) Comparative analyses of the influence of developmental mode on phenotypic diversification rates in shorebirds. Proc R Soc Lond B 273:1619–1624

Venditti C, Meade A, Pagel M (2011) Multiple routes to mammalian diversity. Nature 479:393–396

Yang Z (2006) Computational molecular evolution. Oxford University Press, Oxford, UK

Chapter 18
Phylogenetic Comparative Methods for Studying Clade-Wide Convergence

D. Luke Mahler and Travis Ingram

Abstract A recurring question in ecology and evolutionary biology is whether deterministic evolutionary convergence ever occurs among large sets of species, such as ecological communities or entire evolutionary radiations. Questions about large-scale convergence have featured prominently in discussions of the nature of community assembly and in debates about the relative roles of contingency versus determinism in macroevolution. Until recently, however, there have been relatively few attempts to use a phylogenetic comparative approach to answer questions about clade-level convergence. This is beginning to change with the development of new and more flexible comparative techniques for studying macroevolutionary convergence. In this chapter, we discuss ecological and evolutionary questions that have motivated interest in convergence at large spatial and phylogenetic scales. We review the statistical approaches that have been used to investigate clade-wide convergence, then describe SURFACE, a recently developed method for objectively studying convergence using macroevolutionary adaptive landscape models. We introduce new features within this framework for testing hypotheses about the biogeography of large-scale convergence and for visualizing the relative contributions of different traits to multidimensional convergence, and demonstrate these features using convergent Caribbean *Anolis* lizard faunas. We conclude by discussing the limitations of current approaches for studying clade-wide convergence and highlighting some directions for future research.

D. L. Mahler (✉)
Center for Population Biology, University of California Davis,
2320 Storer Hall, One Shields Avenue, Davis, CA, USA
e-mail: lmahler@ucdavis.edu

T. Ingram (✉)
Department of Zoology, University of Otago, 340 Great King Street, Dunedin, New Zealand
e-mail: travis.ingram@otago.ac.nz

18.1 Why Study Convergent Evolution at the Clade Scale?

The fact that organisms experiencing similar selective pressures often converge by evolving similar adaptations is not controversial, and many studies have elucidated the ecological conditions and genetic and developmental mechanisms responsible for convergence among particular populations or species (Arendt and Reznick 2008; Conway Morris 2003; Losos 2011; Manceau et al. 2010). By contrast, the notion that entire communities or clades could evolve to be deterministically similar remains a source of disagreement among both ecologists and evolutionary biologists (Blondel 1991; Cody and Mooney 1978; Gould 1989; Ricklefs and Miles 1994; Samuels and Drake 1997; Schluter 2000; Schluter and McPhail 1993; Segar et al. 2013).

In ecology, discussions of community convergence have played a role in debates about the mechanisms governing community assembly and structure. Inspired by the perception that similar but historically isolated environments were often host to what appeared to be suites of 'ecologically equivalent' species (e.g., Karr and James 1975; Pianka 1974), a number of workers in the 1970s and 1980s became interested in whether similar ecological settings could lead to the evolution of deterministically similar communities (Cody 1974; Cody and Mooney 1978; Gatz 1979; Karr and James 1975; Keast 1972; Lawton 1984; Melville et al. 2006; Orians and Paine 1983; Orians and Solbrig 1977a; Ricklefs and Miles 1994; Ricklefs and Schluter 1993). Most early community similarity research focused on the non-evolutionary assembly of communities from existing species pools (Cody and Diamond 1975; Fox 1987; Strong et al. 1984), but evidence for similar communities on different continents raised the question of whether matched communities in similar environments could arise through evolutionary processes (Cody and Mooney 1978; Karr and James 1975; Kelt et al. 1996; Ricklefs and Miles 1994; Ricklefs and Travis 1980; Schluter 1986, 1990; Emerson and Gillespie 2008; Segar et al. 2013). Community similarity is most often measured using the phenotypic or functional attributes of species, but related approaches have asked whether communities are similar in other features such as richness and abundance in ecological guilds (Orians and Solbrig 1977b; Segar et al. 2013), the degree of species packing (Gatz 1979; Orians and Solbrig 1977b; Ricklefs and Travis 1980), the axes of variation among species (Wiens 1991; Young et al. 2009), or the relationship between morphology and ecology (Cody and Mooney 1978; Karr and James 1975; Melville et al. 2006; Miles et al. 1987; Montaña and Winemiller 2013; Ricklefs and Miles 1994).

In evolutionary research, the notion of clade-wide convergence has featured most prominently in the debate over the importance of contingency and determinism in macroevolution. Trajectories of evolutionary change are conventionally viewed as being at least somewhat predictable at microevolutionary scales, but idiosyncratic and unpredictable over the longer timescales over which clades diversify. According to this view, commonly referred to as the contingency

hypothesis, both the inevitable differences in the initial conditions for evolution in different lineages and the influence of chance events during the course of evolution are sufficient to preclude highly similar macroevolutionary outcomes (Gould 1989, 2002, 2003; Price et al. 2000; Simpson 1950); (discussed in Beatty 2006, 2008; Erwin 2006; Inkpen and Turner 2012; Pearce 2012; Powell 2009, 2012). Many have challenged this view, citing the apparent emergence of so-called 'replicated adaptive radiations'—independent clades containing similar sets of species that have resulted from diversification in similar environments. With the rise of molecular systematics, phylogenetic investigations of many groups of organisms have revealed patterns of frequent ecological and morphological convergence, often among species that were previously thought to be close relatives based on superficial similarities (Givnish 1997; Losos and Mahler 2010). Such large-scale convergence would provide support for evolutionary determinism and would suggest a more limited role for initial conditions and chance events in determining macroevolutionary outcomes. Phylogenetic patterns suggestive of replicated adaptive radiation have now been reported for cichlid fishes (Clabaut et al. 2007; Kocher et al. 1993; Rüber et al. 1999; Stiassny and Meyer 1999; Young et al. 2009), *Anolis* lizards (Losos et al. 1998; Mahler et al. 2013), Hawaiian spiders (Blackledge and Gillespie 2004; Gillespie 2004, 2005) and plants (Givnish 1999; Givnish et al. 2009), continental radiations of mammals (Madsen et al. 2001; Springer et al. 1997), frogs (Bossuyt and Milinkovitch 2000; Moen and Wiens 2009), damselfishes (Cooper and Westneat 2009; Frédérich et al. 2013), land snails (Chiba 2004), and many other groups (e.g., Alejandrino et al. 2011; De Busschere et al. 2012; Kozak et al. 2009; Patterson and Givnish 2003; Ruedi and Mayer 2001; Ellingson 2013).

Many questions about macroevolutionary convergence remain unanswered. First, because some degree of convergence is expected among any large clades that have diversified from similar ancestors, it is possible that much of the celebrated convergence in replicated radiations is actually unremarkable and not indicative of deterministic evolution (Stayton 2008). Even assuming that replicated radiations are deterministically similar, there has been little investigation into the evolutionary process responsible for such convergence. In particular, while conceptual models of replicated radiation typically involve lineages in different regions being attracted to similar adaptive peaks, methods for inferring the number and position of such peaks on a macroevolutionary landscape have been lacking. In addition, there has been much recent discussion about whether convergence only occurs among members of geographically distinct clades, as might be expected if ecological opportunity and competition regulate the evolution of novel niche specialists (Losos 1996; Wiens et al. 2006), or whether convergence may occur repeatedly within the same region (Kozak et al. 2009; Muschick et al. 2012; Scheffer and van Nes 2006; Ingram and Kai 2014).

Testing hypotheses about processes underlying large-scale convergence has been a challenge, as the long timescales over which clades typically diversify preclude direct observation in nature. As a result, many of our insights into

evolutionary contingency and determinism have come from experimental studies of microorganisms evolving in laboratory settings (Hekstra and Leibler 2012; Kassen 2009; Lenski and Travisano 1994; MacLean 2005; Rainey and Travisano 1998; Saxer et al. 2010; Tyerman et al. 2005) or from computer simulations (Gavrilets and Vose 2005; Pie and Weitz 2005; Scheffer and van Nes 2006; Stayton 2008; Wagenaar and Adami 2004; Yedid and Bell 2002; Yedid et al. 2008). However, in recent years, phylogenetic comparative methods have increasingly been employed to investigate questions about clade-wide convergence in natural systems. To date, most such studies have used phylogenies to identify instances of convergence or to test whether the frequency or fidelity of convergence within or among clades exceeds expectations under simple null models. However, comparative tools that incorporate evolutionary processes into models of convergence on macroevolutionary adaptive landscapes are now available, allowing tests of a wider range of hypotheses about evolutionary convergence using phylogenetic comparative data.

18.2 Historical Development of Methods for Studying Clade-Wide Convergence

Interest in clade-wide convergence predated the advent of modern phylogenetic comparative methods, meaning that the first quantitative techniques for studying convergence among evolutionarily distinct communities were necessarily non-phylogenetic. Starting in the 1970s, community ecologists tested for community convergence by directly comparing the ecological or morphological attributes of species in independent communities (reviewed in Cody and Mooney 1978; Orians and Paine 1983; Blondel 1991; Ricklefs and Miles 1994; Samuels and Drake 1997; Schluter 2000). A common technique for making such comparisons was to construct a multidimensional Euclidean 'morphospace' (Ricklefs and Travis 1980; Wiens 1991; Gatz 1979) and to ask whether the communities showed phenotypic matching (e.g., exceptionally small mean Euclidian distances between species and their nearest neighbors from the other community in morphospace). If the sets of species were well-matched despite their different evolutionary histories, then a case could be made that the communities were convergent (Cody and Mooney 1978; Ricklefs and Travis 1980; Gatz 1979; Ricklefs and Miles 1994). Using this framework, one could test whether communities were more similar than expected using a given null distribution, usually obtained by randomization.

A drawback of this approach is that it is ahistorical and thus does not explicitly test whether organisms have evolved to be more similar to one another than were their ancestors (i.e., whether they are truly convergent or just similar). This point was certainly appreciated by many early workers, and led some to focus on comparisons of communities on different continents (Karr and James 1975; Orians and Solbrig 1977a; Ricklefs and Travis 1980). Schluter (1986) indirectly addressed this problem by comparing the similarity observed among communities in matching

habitats from different regions (and different evolutionary origins, presumably) to the similarity observed among communities in contrasting habitats from the same region. Such a pattern of differences among relatives and similarities with species in similar communities would be consistent with community convergence. However, an alternative explanation for the same pattern is that community matching might be due to ecological sorting from a larger species pool in each region rather than convergent evolution.

A critical step in testing for convergence of communities or radiations was thus the incorporation of evolutionary history into statistical null models. According to this approach, the investigator quantifies the similarity of communities or clades using a standard statistical measure as described above, such as the mean nearest-neighbor distance in morphospace, and then evaluates this statistic against an evolutionary null distribution (Fig. 18.1). We refer to this as the 'statistical' approach in what follows. The null distribution may be generated by simulating trait data on a phylogenetic tree using an evolutionary model that lacks deterministic evolutionary convergence, such as random-walk Brownian motion (BM) (Fig. 18.2) or a single-optimum Ornstein–Uhlenbeck (OU) model. The OU model (Uhlenbeck and Ornstein 1930) includes stochastic evolution as well as attraction toward an 'optimum' trait value, which has the effect of eroding the signal of evolutionary history and reducing the volume of trait space that can be explored (Felsenstein 1988). These features make it useful as a null model that can result in phenotypically similar species without deterministic convergence. The observed measure of convergence can then be compared to this null distribution to test whether putatively convergent groups of species are more similar than expected by chance. A related approach uses phylogenetic simulations to generate a null distribution of a test statistic for ANOVA-like designs (Garland et al. 1993), where the goal is to quantify the phenotypic similarity of unrelated species belonging to categories representing putatively convergent niches (e.g., Brandley et al. 2014; Glor et al. 2003; Harmon et al. 2005; Johnson et al. 2009; Winchester et al. 2014).

Thus far, we have focused on testing for convergence among clades from different regions or among multiple clades that have radiated within a shared region, but another form of 'clade-wide convergence' occurs when multiple sets of lineages in a single large clade converge on one another within a broad geographic region. An alternative method for identifying exceptional convergence in such cases considers how the phenotypic similarity of species is related to the phylogenetic distance between them (Muschick et al. 2012). Stochastic models such as BM predict a general increase in phenotypic distance at increasing evolutionary scales, and despite a high variance in phenotypic distance among distantly related pairs of species, relatively few such pairs are expected to have highly similar phenotypes. In contrast, if convergence is widespread, one can predict an overrepresentation of pairs of phenotypically similar distant relatives. Muschick et al. (2012) devised a test for this pattern that summarizes a plot of phenotypic versus phylogenetic distances using hexagonal binning and tests for a surplus of pairs in bins representing high phylogenetic distance and low phenotypic distance compared to
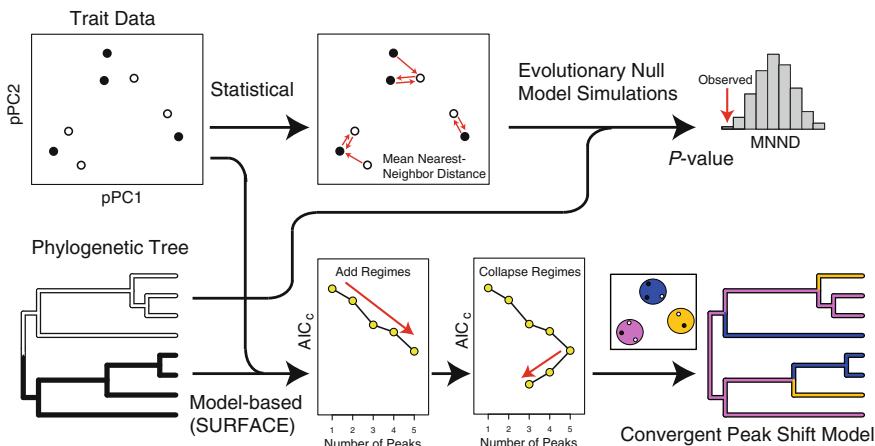
Fig. 18.1 Workflow for two approaches to studying clade-wide convergence using a phylogenetic tree and trait data (multiple phenotypic axes, such as from a phylogenetic principal components analysis). The example data set shown here contains clades in two regions (*white* and *black*), which form three clusters in trait space. The *upper* path illustrates the 'statistical' approach to accounting for phylogeny, in which trait data are used to calculate a statistic summarizing the extent of among-region similarity which is then compared to the same value calculated for data simulated on the tree using an evolutionary null model such as Brownian motion. In this example, the mean of the nearest-neighbor distances between each species and its most similar counterpart in the second region or clade is employed as the summary statistic. The summary statistic can be compared to the null distribution to test for significant clade-wide convergence, but does not incorporate a process model required for model comparison or predictive simulations. The *lower* path shows the approach used by SURFACE to fit macroevolutionary models that incorporate convergence to the tree and trait data. The two stages use stepwise AIC_c first to place peak shifts on the tree and then to identify shifts that are to the same, convergent peak. The method outputs a mapping of the peaks to each branch of the tree and estimates of the peak positions in trait space and the rates of adaptation and stochastic evolution. This 'Hansen' model can then be used in comparison with null or alternative models or to characterize features such as the geographic context (see Fig. 18.3) or dimensionality (see Fig. 18.4) of convergence

evolutionary null models. As long as stasis over long timescales can be ruled out as a cause of similarity between distant relatives, this method provides a means to detect widespread convergence within a clade. While the first application focused on convergence within a region (Lake Tanganyika cichlids), one could also focus on pairs of species from the same or from different regions to test the prevalence of convergence in both geographic contexts.

A drawback of all of these statistical approaches is that they do not explicitly model a process underlying convergent evolution. Instead, they at best compare a measured empirical pattern to a model such as BM that lacks convergent processes, to determine whether the empirical pattern differs from the null expectation in a manner consistent with clade-wide deterministic convergence. While this
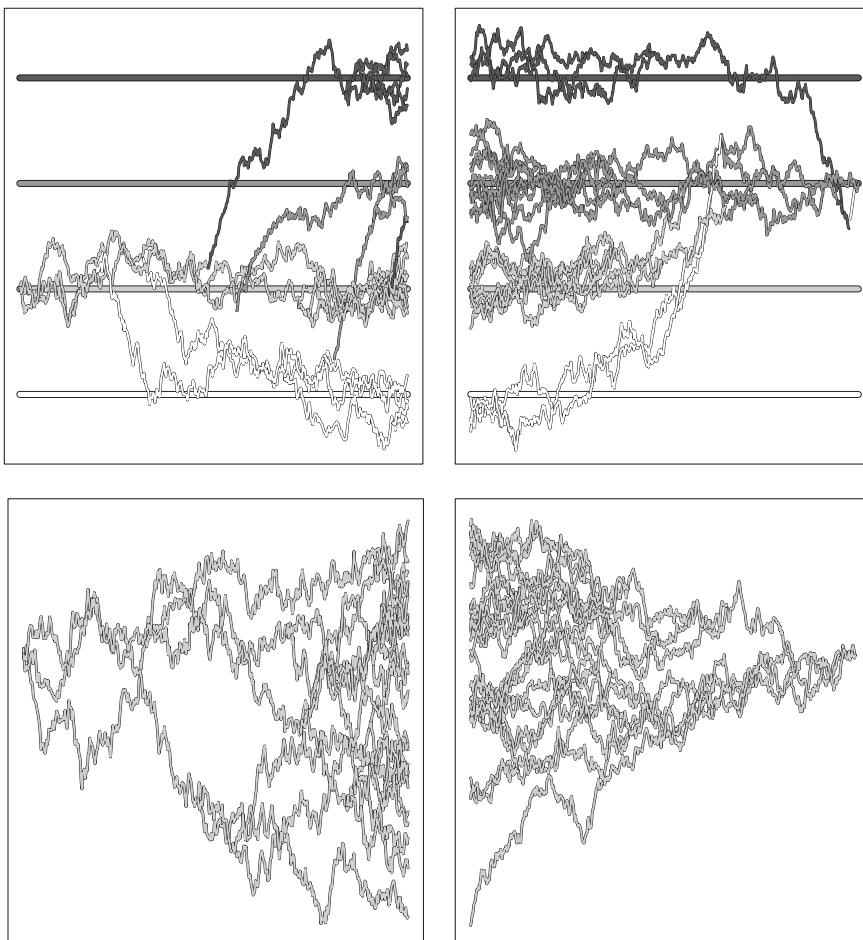
Fig. 18.2 Visualization of deterministic convergence versus chance similarity in two clades. The *top panels* show trait values over time of two clades diversifying to occupy the same set of four adaptive peaks (Ornstein–Uhlenbeck optima, denoted by *horizontal lines*). While the clades differ in ancestral state and the sequence of peak shifts, the resulting faunas are deterministically similar as a result of diversification on the same macroevolutionary landscape. This can be contrasted with the *lower panels*, which show traits evolving under Brownian motion during diversification. In one or a few trait dimensions, some lineages are expected to evolve similar trait values purely by chance; an important goal of any tests for clade-wide convergence is ruling out chance as an explanation for faunal similarity

remains a useful technique for hypothesis testing, it does not provide the means to estimate evolutionary parameters that can result in convergence or allow for comparison with alternative models.

18.3 Development of the Phylogenetic Ornstein–Uhlenbeck Model of the Macroevolutionary Adaptive Landscape and its Utility for the Study of Clade-Wide Convergence

The incorporation of processes that underlie convergence into model-based phylogenetic comparative methods presents new opportunities to gain a mechanistic understanding of replicated radiations. BM does not include the deterministic component necessary to represent adaptive evolution on a macroevolutionary landscape, which makes it useful as a null model but unsuitable if we wish to model convergence explicitly. An important advance was achieved with the realization that versions of the OU model (Uhlenbeck and Ornstein 1930) could be applied to models of adaptive evolution (Chaps. 14 and 15). The deterministic component of the OU process causes a species' mean trait value to be pulled toward an optimum, with a force proportional to its distance from the optimum (Chap. 15). Lande (1976) showed that this model can describe the evolution of a continuous trait in a population under a balance of stochastic mutation and genetic drift and deterministic selection that is directional when the population is approaching the optimum and stabilizing when it is at the optimum. While the parameters of the OU model estimated from comparative data often do not correspond well to population genetic parameters (Harmon et al. 2010), the model has proven to be a useful representation of adaptive evolution, in which the optimum can be interpreted as the location of an adaptive peak (Felsenstein 1988; Martins 1994).

An important advance came when Hansen (1997) demonstrated how the OU model could be used to model peak shifts on a macroevolutionary adaptive landscape (e.g., Simpson 1944). If certain lineages within a clade begin evolving toward a new peak, such as equids shifting from browsing to grazing, this can be modeled as these lineages becoming attracted to a new optimum trait value. Hansen (1997) showed how this peak-shift model could be fit to a phylogenetic tree and trait data, and Butler and King (2004) subsequently introduced the 'OUCH' methodology as a generalized framework for fitting such 'Hansen' models by 'painting' the branches of a tree with multiple selective regimes, each corresponding to a hypothesized peak. The multipeak OU model can thus be interpreted as a representation of the macroevolutionary landscape (Simpson 1944) and can be used to test hypotheses such as whether clades in multiple regions have reached the same set of peaks (Fig. 18.2). The parameters of the model are the positions of one or more optima θ; the rate α at which species adapt toward their present optimum; and the Brownian rate parameter σ^2 that governs the magnitude of stochastic fluctuations in trait values (an additional parameter represents the root state and is typically assumed to come from the stationary distribution of the OU process rather than being estimated—see Chap. 15 for additional discussion). The macroevolutionary landscape is characterized by one or more adaptive peaks, with the relative rates of adaptation and stochastic evolution roughly indicative of the steepness of the peaks. We deal here with 'Simpsonian' macroevolutionary

landscapes that measure individual fitness as a function of phenotypic values, in contrast to 'Wrightian' landscapes that measure population mean genotype fitnesses (Hansen 2012; Svensson and Calsbeek 2012).

The Hansen model has been widely used to test adaptive hypotheses, such as whether species in different regimes (typically defined based on ecological contexts such as discrete habitats) repeatedly evolve toward different phenotypic optima. Other elaborations of the OU model allow variation among regimes in α and σ^2 in addition to the positions of optima (Beaulieu et al. 2012; Chap. 15), increasing the variety of hypotheses that can be tested. The Hansen model is a valuable tool for comparative hypothesis testing, but there is an important caveat that limits its applicability to testing for replicated radiation. The method requires an a priori adaptive hypothesis, such that one must paint the hypothesized regimes onto specific branches to reflect taxa already thought to be convergent in advance of testing for clade-wide convergence. This approach is suitable for testing certain hypotheses about convergent evolution, such as whether particular ecological or behavioral shifts consistently lead to adaptation of similar morphological traits (e.g., Collar et al. 2011; Frédérich et al. 2013; Lapiedra et al. 2013). However, if we wish to assess the extent of convergence using phenotypic data alone, it would be circular to first use phenotypic similarity to assign species to regimes. To deal with this limitation, we devised an algorithm for fitting Hansen models to a data set in the absence of a priori hypotheses and thus objectively identifying convergent peak shifts (Ingram and Mahler 2013).

18.4 Using SURFACE to Infer a Macroevolutionary Adaptive Landscape from Comparative Data

As we have described, tests for clade-wide convergent evolution have been dominated by two approaches. One statistically quantifies the similarity of communities or faunas and uses null model comparisons to assess whether the observed similarity exceeds what would be expected by chance. This method has the benefit of objectivity, but does not incorporate a mechanism underlying convergence. The second approach uses model-based comparative methods to test whether independent shifts to shared selective regimes result in convergent phenotypic evolution. This method incorporates adaptive processes, but its reliance on potentially subjective regime specification is problematic if we wish to test for clade-wide phenotypic convergence without using independent data to define regimes.

To bridge this gap, we introduced the SURFACE method, which constructs a representation of the macroevolutionary adaptive landscape, taking as inputs only continuous trait data and a tree (Ingram and Mahler 2013). SURFACE (a recursive acronym for 'SURFACE Uses Regime Fitting and AIC to model Convergent Evolution') uses stepwise AIC (Alfaro et al. 2009; Thomas and Freckleton 2012) to first identify peak shifts well supported by the data and then to identify whether any

of these shifts involve convergence toward the same peaks (Fig. 18.1). The method is implemented as the R package 'surface', which contains functions for running the analysis, simulating data sets, and visualizing the results.

Here, we briefly outline the steps involved in a SURFACE analysis, and refer the reader to Ingram and Mahler (2013) for a detailed description. First, a single-peak OU model is fit to the phylogeny and continuous trait data using maximum likelihood. This approach can handle multidimensional trait data by making the simplifying assumption that there are no covariances between traits in the rates of adaptation (α) or stochastic evolution (σ^2), which is most likely to be valid if the traits are orthogonal axes such as those obtained from a principal components analysis. This assumption allows the log likelihoods calculated separately for each trait to be added together to give the overall model log likelihood L. A set of all candidate models in which a peak shift is added to one branch in the tree is then generated (in the present implementation, each branch may experience only a single peak shift, which occurs at its origin). Each model is fit, and log likelihoods are added across traits as before. For each candidate model, the small sample size-corrected Akaike Information Criterion (AIC_c) is calculated as a measure of model performance that accounts for the model complexity (number of parameters, p) and sample size n (number of species k multiplied by number of traits m).

$$AIC_c = -2 \log L + 2k + \frac{2p(p+1)}{n-p-1} \qquad (18.1)$$

The addition of one peak shift adds one parameter per trait to account for the new estimated optima and one parameter to represent the phylogenetic placement of the peak shift. The candidate model that most improves (i.e., reduces) the AIC_c is selected, a peak shift is placed at the origin of the corresponding branch, and the process is iterated to place additional peak shifts until the AIC_c ceases to improve.

This 'forward' phase of the method yields a Hansen model containing some number of peak shifts, each of which is toward a different peak. The second, 'backward' phase evaluates the fit of models in which sets of shifts are toward to the same convergent peak. Reducing the number of peaks (and optima) in the model may improve the AIC_c if the model likelihood remains high despite the reduction in model complexity. This second stepwise process is iterated until model improvement stops and a final model is identified. The extent of convergence in this model can be quantified in a number of ways, including the reduction in the number of peaks in the backward phase and the number of peak shifts that are toward convergent peaks (see 18.5). Once quantified, measures of convergence from the fitted model can be compared to a null distribution obtained by running SURFACE on data simulated under BM or other models that lack deterministic convergence. Additionally, parametric bootstrapping can be carried out to construct approximate confidence intervals on the convergence parameters or other parameters of interest, by running SURFACE on many data sets simulated under the fitted Hansen model. These approaches allow the researcher to evaluate whether the extent of convergence in a model is greater than expected by chance.

18.5   Extending SURFACE to Ask Questions About the Nature of Convergence in Replicated Radiations

SURFACE provides a Hansen model representing the macroevolutionary adaptive landscape, as well as several measures of the extent and frequency of convergence, but additional steps are needed to interpret the details of any detected convergence. Here, we describe new methods for characterizing the biogeographic pattern of convergence and for comparing the extent of convergence among traits.

   a. The geographic context of convergence

   Biogeography is central to many hypotheses about clade-scale convergence and must be incorporated if we wish to use SURFACE to infer whether convergence tends to occur because radiations in different regions are replicated (e.g., Chiba 2004; Mahler et al. 2013; Ellingson 2013) or because radiations generate locally replicated adaptive diversity within regions (e.g., Kozak et al. 2009; Muschick et al. 2012). Some degree of replicated radiation can be inferred if the same peaks are occupied by lineages in each region, and such an inference may be straightforward in cases where each region is occupied by a distinct subclade. However, if the ancestral geographic locations of lineages are uncertain, we need appropriate statistical methods to account for this uncertainty in evaluating whether adaptive peak shifts occurred in the same or in different regions.

   We describe a simple approach to inferring the biogeography of convergence using a Hansen model fitted by SURFACE, which contains some number of peak shifts, each assigned to a branch in the tree. We combine this information with hypothesized biogeographic histories that include the timing of dispersal events between regions and can be used to estimate the region in which each peak shift occurred. The calculations described below require that each node in the tree can be assigned to a single geographic region; if some ancestors are though to span multiple regions, a reasonable approach may be to treat this as uncertainty and sample histories that include the different regions. Biogeographic history estimates may come from a variety of techniques including ancestral character reconstruction using likelihood or parsimony and stochastic character mapping (Huelsenbeck et al. 2003). While these histories can include one or more dispersal events along a branch, we consider only the geography at the beginning and at the end of each branch. We do this so that geography is defined at the same resolution as peak shifts, as SURFACE is limited to inferring one peak shift per branch. We make the assumption that the region occupied at the end of a branch (rather than the region occupied by the parent species when the branch originated) is the region in which any peak shift occurred. Unless specified otherwise, the peak shift placed at the root of the tree in the Hansen model is assumed not to correspond to a geographic shift.

   To illustrate our approach, we incorporate estimates of geographic history into an analysis of ecomorphological convergence in Greater Antillean anoles (Fig. 18.3). Mahler et al. (2013) used SURFACE to show that there are many more instances of convergence in this group than expected by chance, and inferred that
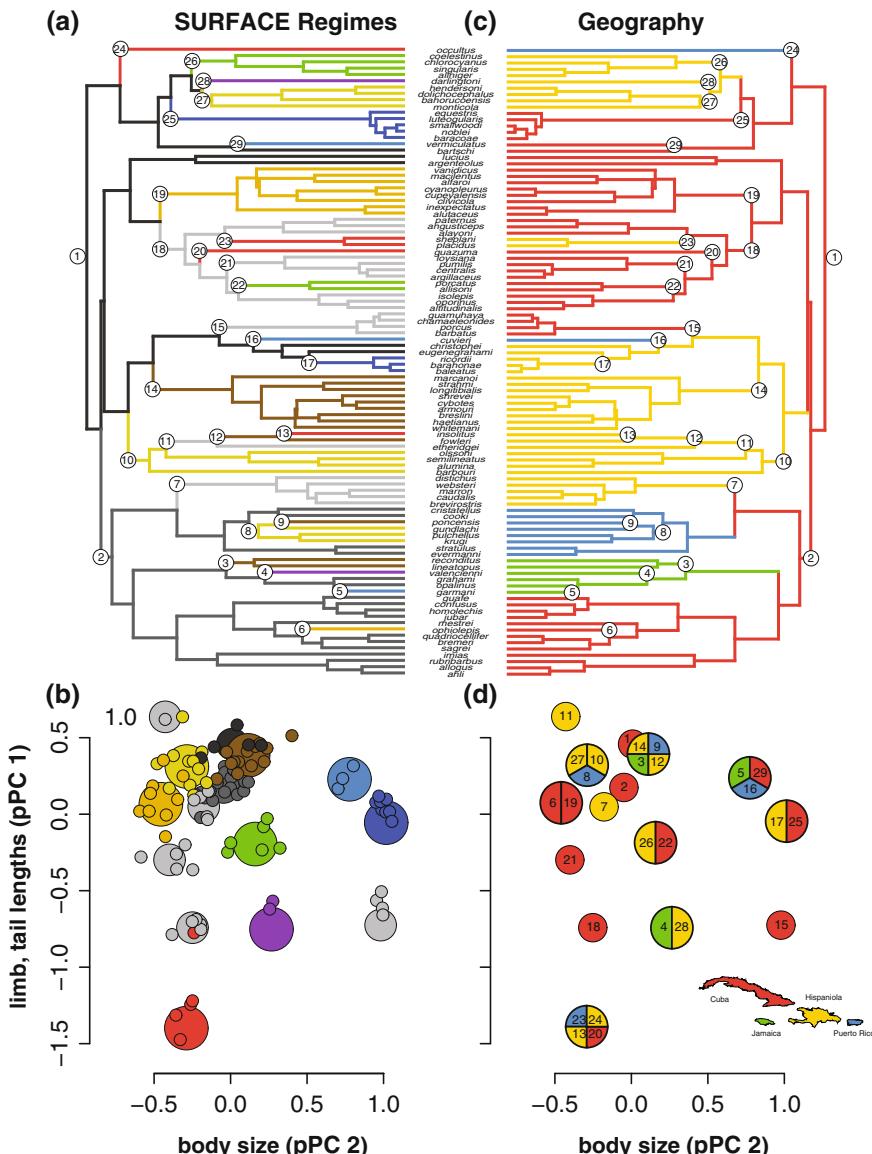
Fig. 18.3 Illustration of a Hansen model obtained using SURFACE that illustrates the biogeography of exceptional clade-wide morphological convergence in Greater Antillean *Anolis*. The *left panels*, adapted from (Mahler et al. 2013), illustrate the placement of adaptive peaks on the branches of the phylogenetic tree (**a**) and in morphospace (**b**). Branches in color indicate convergent peaks (that attract more than one lineage), while non-convergent peaks are in gray scale. **b** shows the first two dimensions of a four-dimensional morphospace for Greater Antillean anoles derived from a phylogenetic principal components analysis (pPCA). *Large circles* denote estimated positions for peaks on the macroevolutionary landscape (*circles* are slightly *larger* for convergent peaks than non-convergent ones), and small circles denote trait values for the 100

extant anole species in the data set. Colors of species' trait values and the optima to which they are attracted correspond to **a**. The *right panels* show the biogeography of anoles on the tree (**c**) and of the peaks in morphospace (**d**). The phylogeny in **c** is identical to **a**, but branches are colored according to the Greater Antillean island occupied at the end of the branch (see text) in one stochastic character map estimate of *Anolis* geographic history. *Panel* **d** depicts the estimated positions of adaptive peaks in morphospace as in **b**, but colored by island to indicate the geography of convergent and non-convergent peak shifts. Numbers indicate which lineages from **c** have shifted to each adaptive peak, allowing assessment of whether shifts to convergent peaks occurred within the same geographic region (e.g., shifts 6 and 19), in different regions (e.g., shifts 17 and 25) or both (e.g., shifts 8, 10 and 27). Note that for the morphospace panels, peaks that appear to overlap are distinct in other trait dimensions (not illustrated) that were part of the SURFACE analysis

most instances of convergence involved lineages from multiple islands. To formalize the latter result and to uncover additional details about the geography of Greater Antillean anole convergence, we use stochastic character mapping with the 'make.simmap' function in the 'phytools' R package (Revell 2012; Chap. 4) to generate 1,000 biogeographic histories of transitions between the four islands (Cuba, Hispaniola, Jamaica, and Puerto Rico; Fig. 18.3c). We then reconsider the macroevolutionary landscape inferred by SURFACE (Mahler et al. 2013) in an explicitly geographic context.

Having inferred the phylogenetic positions of both peak shifts and geographic shifts, we can ask a variety of questions about the geography of convergence. The biogeographic question most relevant to tests for replicated radiation is whether convergence to a shared peak typically occurs in distinct regions or within a single region. To assess this, we examine two alternative measures that categorize convergence by geography: one uses the geography of individual convergent peak shifts, while the other uses the geography of pairs of convergent lineages. While these measures provide similar information, they allow us to ask somewhat different questions about the geography of convergence.

The first approach classifies each peak shift based on the geographic context of the lineages that have reached that peak. We first count the number of peak shifts in the model fitted by SURFACE that are toward convergent peaks, denoted as c (Ingram and Mahler 2013). For each of the c convergent peak shifts, we ask whether the shift occurred toward a peak occupied only by lineages that occur in different regions from that of the focal lineage, and we count the total number of such shifts as $c_{between}$ (e.g., shifts 3, 8 and 29, among others, in Fig. 18.3d). Next, we count the number of convergent shifts that occur toward a peak that is only occupied by lineages from the same region, c_{within} (e.g., shifts 6 and 19 in Fig. 18.3d). Finally, we count the number of shifts to peaks occupied both by other lineages from the same region, as well as lineages from different regions, c_{both} (e.g., shifts 10, 12, and 27, among others, in Fig. 18.3d). These three geographic convergent shift statistics sum to c. Note that the total number of convergent shifts to peaks occupied in more than one region is the sum of $c_{between}$ and c_{both} and that the total number of geographically replicated convergent shifts (i.e., shifts to peaks occupied by at least one other lineage from the same region) is the sum of c_{within}

and c_{both}. For *Anolis*, we averaged these calculations across the 1,000 stochastically mapped geographic histories. As expected, most of the convergent peak shifts (91 %) were to peaks occupied on multiple islands (i.e., $c_{between} + c_{both}$), while very few (9 %) were to single-island peaks (i.e., c_{within}; Fig. 18.3d). However, it was not uncommon for independent lineages from the same island to adapt to the same adaptive peak, and 36 % of shifts were to peaks occupied by at least one other lineage from the same island (i.e., $c_{within} + c_{both}$; Fig. 18.3d).

The second approach quantifies the geography of pairwise cases of convergence, examining all pairs of peak shifts that were toward the same peak. We define the number of such cases as cc: a peak reached by two shifts represents a single case of convergence; a peak reached by three shifts represents three pairwise cases, and so on. For each of the cc pairwise cases of convergence, we identify whether the two shifts occurred in the same region (e.g., for convergent peaks in Fig. 18.3d, whether pairs of shifts have the same color) or in different regions (pairs of shifts have different colors in Fig. 18.3d). We add these values to estimate the number of within-region cases of pairwise convergence (cc_{within}) and the number of between-region cases ($cc_{between}$), which sum to cc (Ingram and Kai 2014). The alternative measures c and cc provide similar information and will differ more in cases when many lineages converge on a small number of peaks. In anoles, we found that 82 % of pairwise cases of convergence occurred between islands (Fig. 18.3d), confirming the replicated nature of Greater Antillean anole radiations.

Specific hypotheses about replicated radiation can be carried out by comparing any of these geographic measures of convergence to null distributions, generated either by randomizing the positions of geographic and/or peak shifts or by analyzing simulated data sets with SURFACE. Related approaches could be used to ask additional biogeographic questions, such as whether convergence is more common in some regions than in others, whether shifts toward certain peaks only occur in certain types of regions (e.g., large versus small; temperate versus tropical), and whether peak shifts coincide with the colonization of novel areas. In anoles, we found that on average, 20 % of peak shifts occurred on branches containing geographic shifts, while 56 % of geographic shifts occurred on branches containing peak shifts. This indicates that most peak shifts occur within islands, but that about half of the anole colonizations to new islands nonetheless coincided with an adaptive peak shift. In addition to testing new hypotheses, there is scope for extending the methods described here to accommodate common biogeographic scenarios with added complexity, such as species that span multiple regions or regions that do not have discrete boundaries.

   b. Partitioning the signal of convergence among traits

Another important feature of clade-wide convergence is the degree to which convergent evolution varies among traits in a multidimensional data set. The SURFACE algorithm combines evidence across multiple traits when identifying the best-fitting model at each step, making the simplifying assumption that the evolution of each trait is governed by independent parameters (α and σ^2).

However, a signal of convergence may arise in different ways: each trait may contribute approximately equally to the model improvement, or one or more traits may show very strong convergence, while others show little or none. In such a case, the additional non-convergent traits may reduce our ability to detect the strong signal of convergence in other phenotypic dimensions, and we may draw false conclusions about the dimensionality of convergence.

Here, we describe how one can visualize the contribution of each trait to the improvement in the overall model support (i.e., AIC_c) during the course of model selection by SURFACE. We have previously visualized the change in AIC_c throughout the forward and backward phases using a line graph of model AIC_c against the number of peaks at each step (Ingram and Mahler 2013; Mahler et al. 2013). In a data set with evidence for many convergent peak shifts, this graph will appear roughly triangular, with an AIC_c decrease from left to right as peaks are added during the forward phase, and a second decrease from right to left as convergent peaks are collapsed during the backward phase.

To divide the AIC_c among traits, we must consider both components of the AIC_c (Eq. 18.1). The fit of the model to each trait, captured as the deviance $(-2 \log L)$, can easily be partitioned among traits as it is based on the sum of trait-specific log likelihoods. The second component of the formula is the 'penalty' term, which captures the complexity of the model (number of parameters p) and the correction for finite sample size (n). This component is not trait-specific (because the parameters representing the phylogenetic positions of shifts are shared across traits), but for the purpose of visualization, we divide this term by the number of traits (m) and then add this value to the deviance for each trait i to obtain a 'partial AIC_c':

$$\text{partial } AIC_c(i) = -2logL_i + \frac{1}{m}\left(2p + \frac{2p(p+1)}{n-p-1}\right). \tag{18.2}$$

We standardize the overall AIC_c and the partial AIC_c of each trait to initial values of zero and then plot the partial AIC_c values along with the overall AIC_c as a function of the number of peaks through both phases of the analysis. For each point on the line graph, the sum of the partial AIC_c values is equal to the overall AIC_c.

The extent to which the partial AIC_c for each trait declines gives an indication of how much it contributes to the addition of peak shifts during the forward phase and to the identification of convergent peak shifts during the backward phase. If all traits contribute roughly equally, their lines will each show a similar decline, while if the signal is dominated by a single trait, the latter's partial AIC_c will decline substantially, while those of the other traits may decline little or even increase. This visualization can be done using the function 'surfaceAICPlot' in the 'surface' package, which has an option *traitplot* that can be set to *dev* for deviance or *aic* to use the partial AIC_c values. If the deviance is used, the values will sum to give the total model deviance rather than the AIC_c; as the deviance cannot improve as the number of parameters decreases during the backward phase, we find the visualization using partial AIC_c more intuitive.

To illustrate this approach, we show a partial AIC$_c$ plot from the SURFACE analysis of ecomorphological convergence in the anole data set, which consists of four traits derived from a phylogenetic principal components analysis (Revell 2009; Fig. 18.4). This visualization shows that the first two trait axes, which correspond to relative limb length and body size and combine to explain 73 % of the total variation, each contribute strongly to the overall improvement in model fit. The third axis, which loads with relative tail and limb lengths, showed moderate improvement, while the AIC$_c$ for the fourth axis, representing toepad lamella number, worsened throughout the forward phase. This trait-by-trait examination of the anole analysis reinforces the longstanding view that convergence in anoles is multidimensional (e.g., Harmon et al. 2005), but indicates that not all trait axes show the same signal. While repeating a SURFACE analysis following the post hoc removal of traits that failed to contribute to model improvement constitutes data dredging and would thus be problematic for hypothesis testing, care should be taken at the outset not to include traits that lack biological relevance (such as very minor axes of a principal components analysis) or that are unrelated to the ecological and evolutionary questions of interest.

18.6 Caveats, Future Directions, and Conclusions

18.6.1 Caveats

Direct inference of convergence requires that the similarity of putatively convergent species can be compared to that of their ancestors. Because empirical information about ancestral phenotypes is typically lacking, detecting convergence using comparative data is a challenge. By explicitly modeling processes that can cause convergence and examining the fit of such models to empirical data, we can use phylogenetic approaches to gain new insights into convergence. Nonetheless, such approaches have only recently been developed, and current methods exhibit several limitations. Here, we discuss caveats associated with the use of existing landscape models and propose several suggestions for the improvement and further development of comparative tools for studying convergence.

The SURFACE method generates an approximation of the macroevolutionary adaptive landscape in the form of a multiple-peak phylogenetic Hansen model. While this model incorporates important evolutionary processes, it is a rather simple representation of the landscape (Hansen 2012; Ingram and Mahler 2013). The Hansen model estimated by SURFACE contains estimates of the phylogenetic positions of adaptive peak shifts, the positions of the peaks in morphospace (θ), the trait-specific rate of adaptation of lineages toward those peaks (α), and the trait-specific rate of stochastic evolution (σ^2). A more general model of evolution on the macroevolutionary adaptive landscape would permit variation in the evolutionary parameters, so that peaks might vary in height or steepness or some lineages might
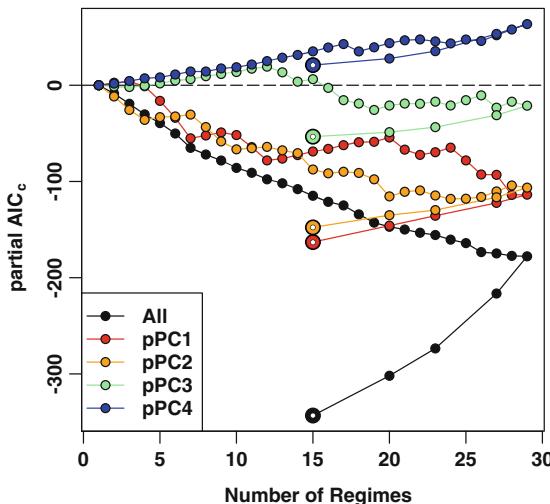
Fig. 18.4 Visualizing the relative contributions of individual traits to multidimensional convergence. This figure shows the relative 'partial AIC$_c$' scores for each trait axis (*colored sequences*) as well as the relative multidimensional AIC$_c$ scores (*black sequence*) for all of the models fit during a single SURFACE analysis of Greater Antillean anoles. For each trait, partial AIC$_c$ values are standardized relative to the partial AIC$_c$ score of the initial single-peak Hansen model. Starting at the left with a partial AIC$_c$ value of zero (corresponding to the single-peak OU model), each sequence records that trait's relative model support for each Hansen model (y-axis) as the number of independent adaptive peaks (x-axis) is increased during the forward SURFACE phase and then reduced during the backward phase as convergence among independent lineages is permitted. Although cumulative model support increases at each SURFACE step (decreasing AIC$_c$ values at each step in the full model sequence), model support for individual traits does not improve uniformly as peaks are added, and peak shifts are sometimes added even if they decrease model support for some traits, provided that these losses are offset by gains in model support for other traits. An estimate of the relative contribution of each trait to the fit of the final Hansen model may be obtained by comparing the partial AIC$_c$ values of all traits for this model (*larger circles* with *white dots* in the center). In the case of Greater Antillean *Anolis*, the overall fit of the final Hansen model is driven in large part by relative limb length (pPC1) and body size (pPC2), with a smaller contribution from relative tail and forelimb lengths (which have contrasting loadings on pPC3). The final model actually provides an inferior fit to toepad lamella data (pPC4) than the initial single-peak model. Note that while individual traits varied in whether they supported the addition of new peaks during the forward phase, all traits supported the simplification of the macroevolutionary landscape as similar peaks were modeled as convergent

have higher intrinsic rates of adaptation or stochastic evolution. The flexible Ornstein–Uhlenbeck 'OUwie' model described by Beaulieu et al. (2012; Chap. 15) permits α and/or σ^2 to vary across the tree, but is not presently implemented to operate in the absence of an a priori hypothesis about the phylogenetic positions of shifts in $\boldsymbol{\theta}$, α, and σ^2. While it is in theory straightforward to use stepwise model selection to compare flexible Ornstein–Uhlenbeck models, it may prove challenging in practice due to the need to fit large numbers of candidate models at each step, and the possibility that many data sets will not contain enough information to

distinguish among alternative plausible parameter combinations for such complex models (Beaulieu et al. 2012).

SURFACE also assumes that the positions of peaks on the adaptive landscape are static, but both Simpson's description (Simpson 1944) and recent elaborations (Arnold et al. 2001; Hansen 2012) allow dynamic macroevolutionary landscapes on which species adapt toward peaks whose positions shift over time. If the landscape is dynamic, our ability to identify and interpret clade-wide convergence will likely depend on whether peaks move in synchrony in different regions (e.g., if they are tracking shared climatic changes) or more idiosyncratically. Recent methodology allows the incorporation of moving optima into the inference of OU model parameters (Bartoszek et al. 2012; Hansen et al. 2008), and extensions of these methods may allow the inference of convergence even in cases where the adaptive landscape is dynamic. Another possibility is to relax the assumption of SURFACE that each trait is evolutionarily independent in terms of its α and σ^2 and to evaluate the influence of correlated evolution as lineages converge to shared adaptive peaks (Bartoszek et al. 2012).

As with any statistical tool, users of SURFACE should ensure both that they have a data set that is sufficiently large and complete to support the fitting of complex models (see discussion in Ingram and Mahler 2013) and that the assumptions of the standard Hansen model yield realistic parameter estimates for their data. The examination of the estimated positions of trait optima in morphospace may aid in evaluating the appropriateness of the model, with two possible interpretations in the event that any estimated optima are extreme, falling far outside the range of trait values in morphospace. First, the assumption that all evolutionary regimes have the same rates of adaptation and stochastic Brownian evolution may be invalid, and the optimum might be estimated to be distant because the rate of adaptation to the peak is constrained to equal the lower rate supported elsewhere in the clade. Alternatively, the poorly matched species may in fact be experiencing directional selection toward a distant optimum (though the position of the optimum itself is unlikely to be estimated well). In addition to visual inspection, posterior predictive simulation may be useful for checking whether empirical measures of convergence indeed arise from the fitted model (e.g., see Mahler et al. 2013). Also, pairing a SURFACE analysis with a 'statistical' approach for testing for convergence as described above may help to confirm the robustness of the pattern of similarity implied by the model fit.

Because a Hansen model estimated by SURFACE contains a mix of parameters representing the evolutionary model (θ, α, and σ^2) and parameters describing the extent of convergence in this model (e.g., c, cc, and the geographic variables discussed in 18.5), uncertainty in the model must be considered at multiple levels. A single SURFACE analysis does not provide an estimate of this uncertainty, but it is straightforward to calculate confidence intervals for the rates and optima using parametric bootstrapping, simulating many data sets under the fitted model and re-estimating the parameters. A similar bootstrapping approach can be used to infer confidence intervals for measures of convergence, although this requires the

more time-consuming process of simulating under the fitted model and running SURFACE on each resulting data set. This approach can in theory be extended to incorporate uncertainty in the topology and branch lengths of the phylogeny itself, by running analyses on a sample of trees (e.g., from the posterior distribution from a Bayesian phylogenetic analysis). While variation in topology among trees precludes direct comparisons of the Hansen models returned by SURFACE, one can combine bootstrapped estimates of convergence parameters across trees to obtain confidence intervals that more completely account for uncertainty.

There are many possible Hansen models that could be fit to a given data set, and while the stepwise algorithm used in SURFACE provides a means of navigating among candidate models, it also results in many models never being evaluated. In some cases, the early fixation of well-supported shifts may preclude the later discovery of globally superior Hansen models that do not include those shifts. Unlike with the model parameters, it is not straightforward to measure the uncertainty in the fitted model that is due to these stepwise constraints. A partial solution to this problem is provided in the *sample_shifts* option of the 'surfaceForward' and 'runSurface' functions in the 'surface' package, which permits the fixation of suboptimal shifts (chosen randomly from a sample of models above a specified support threshold) during an individual SURFACE run. By repeatedly running SURFACE using relaxed model selection criteria, it is possible to obtain a set of models for which an important element of path dependency has been relaxed (see Mahler et al. 2013 for an example). Although such a set of models is not statistically analogous to a sample of models from a Bayesian posterior distribution, it can nonetheless be used to heuristically assess the influence of path dependency on parameter estimates, as well as to potentially identify models that may be superior to the model returned by the standard SURFACE analysis.

In the future, it may be possible to simultaneously address several of these issues by exploring model and parameter space in a Bayesian framework. Bayesian Markov Chain Monte Carlo methods have recently been employed for a similar comparative purpose—identifying the locations of phylogenetic shifts in the evolutionary rate without a priori information (Eastman et al. 2011; Revell et al. 2012; Venditti et al. 2011), and a similar approach could be employed to model adaptive peak shifts. To identify shifts, it would be necessary not only to estimate Hansen model parameters, but also to estimate the numbers of total and convergent peak shifts, meaning that the algorithm would need to evaluate models that differ in the number of estimated parameters. Eastman et al. (2011) used reversible-jump MCMC to sample from models of varying complexity to determine the number and placement of evolutionary rate shifts that have occurred in a clade, and this approach may be suitable for examining shifts among evolutionary regimes. Alternatively, the clustering of convergent lineages into discrete regimes in morphospace might be modeled using a Dirichlet process prior (e.g., Heath et al. 2012) in which both the number and phylogenetic branch composition of regimes, as well as the evolutionary parameters that characterize these regimes are estimated.

18.6.2 Future Directions

The development of methods for explicitly modeling macroevolutionary convergence provides a powerful framework for studying convergence among entire communities or clades. This is leading to renewed investigation of key questions in ecology and evolution, such as whether radiations in different biogeographic regions are more similar than expected by chance and whether significant convergence occurs within single regions. In combination with suitable 'statistical' approaches to measuring the pattern of similarity, macroevolutionary models of convergence provide the means to ask many additional questions about phenotypic convergence.

One potentially useful application of SURFACE might be to directly model convergence in species' ecological attributes (e.g., continuous measures of habitat use, climatic niche preferences, or diet), rather than morphological traits that are thought to reflect ecological adaptations. The Hansen model fit to the ecological attributes of extant species might reveal whether distinct ecological niches are stable and convergent over long timescales. Further, the evolutionary correspondence between ecology and morphology could be investigated explicitly through comparison of macroevolutionary adaptive landscapes separately estimated for clades using ecological versus morphological attributes. By examining the positions of peak shifts in each Hansen model, one could use such information to ask whether species adopt novel ecological preferences prior to evolving morphological specializations, or alternatively whether morphological innovations precede novel resource use. Likewise, a comparison of ecological and morphological macroevolutionary landscapes might complement functional studies to reveal 'many-to-one mapping,' an alternative to morphological convergence in which lineages adapting to similar ecological pressures evolve different morphological solutions (Alfaro et al. 2005; Bock 1980; Bock and Miller 1959; Losos 2011).

Models of the macroevolutionary landscape may also be useful for answering longstanding questions about the sequence of adaptations during the evolutionary assembly of communities and clades. Hansen models provide an estimate of the temporal sequence of peak shifts, though we note that accurate placement of peak shifts on the branches of the phylogeny becomes more difficult for deep branches or when peak shifts are frequent (Ingram and Mahler 2013). Information about the timing of peak shifts could be used to test whether adaptive peaks in replicated radiations are colonized in the same sequence or whether the evolutionary assembly of similar radiations is more idiosyncratic (Ackerly et al. 2006; Losos et al. 1998; Sallan and Friedman 2012; Streelman and Danley 2003); reviewed in (Glor 2010). Other questions about the filling of morphospace that could be addressed are whether peak shifts tend to occur between nearby peaks (versus large jumps through morphospace) and whether peaks at the periphery of morphospace tend to be discovered later in the course of adaptive radiation (Gavrilets and Vose 2005; Price 1997; Ricklefs and Travis 1980). The continued application and elaboration of macroevolutionary adaptive landscape models has the potential to address many exciting questions about convergent evolution in adaptively radiating clades.

References

Ackerly DD, Schwilk DW, Webb CO (2006) Niche evolution and adaptive radiation: testing the order of trait divergence. Ecology 87:S50–S61

Alejandrino A, Puslednik L, Serb JM (2011) Convergent and parallel evolution in life habit of the scallops (Bivalvia: Pectinidae). BMC Evol Biol 11:164. doi:10.1186/1471-2148-11-164

Alfaro ME, Bolnick DI, Wainwright PC (2005) Evolutionary consequences of many-to-one mapping of jaw morphology to mechanics in labrid fishes. Am Nat 165:140–154. doi:10.1086/429564

Alfaro ME, Santini F, Brock C, Alamillo H, Dornburg A, Rabosky DL, Carnevale G, Harmon LJ (2009) Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. Proc Natl Acad Sci 106:13410–13414. doi:10.1073/pnas.0811087106

Arendt J, Reznick D (2008) Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation? Trends Ecol Evol 23:26–32. doi:10.1016/j.tree.2007.09.011

Arnold SJ, Pfrender ME, Jones AG (2001) The adaptive landscape as a conceptual bridge between micro- and macroevolution. Genetica 112–113:9–32

Bartoszek K, Pienaar J, Mostad P, Andersson S, Hansen TF (2012) A phylogenetic comparative method for studying multivariate adaptation. J Theor Biol 314:204–215. doi:10.1016/j.jtbi.2012.08.005

Beatty J (2006) Replaying life's tape. J Philos 103:336–362

Beatty J (2008) Chance variation and evolutionary contingency: Darwin, Simpson (The Simpsons) and Gould. The Oxford Handbook of Philosophy of Biology. Oxford University Press, Oxford

Beaulieu JM, Jhwueng D-C, Boettiger C, O'Meara BC (2012) Modeling stabilizing selection: expanding the Ornstein–Uhlenbeck model of adaptive evolution. Evolution 66:2369–2383. doi:10.1111/j.1558-5646.2012.01619.x

Blackledge TA, Gillespie RG (2004) Convergent evolution of behavior in an adaptive radiation of Hawaiian web-building spiders. Proc Natl Acad Sci 101:16228–16233

Blondel J (1991) Assessing convergence at the community-wide level. Trends Ecol Evol 6:271–272

Bock WJ (1980) The definition and recognition of biological adaptation. Am Zool 20:217–227

Bock WJ, Miller WD (1959) The scansorial foot of the woodpeckers, with comments on the evolution of perching and climbing feet in birds. Am Mus Novit 1931:1–45

Bossuyt F, Milinkovitch MC (2000) Convergent adaptive radiations in Madagascan and Asian ranid frogs reveal covariation between larval and adult traits. Proc Natl Acad Sci 97:6585–6590

Brandley MC, Kuriyama T, Hasegawa M (2014) Snake and bird predation drive the repeated convergent evolution of correlated life history traits and phenotype in the Izu Island scincid lizard (Plestiodon latiscutatus). PLoS ONE 9:e92233. doi:10.1371/journal.pone.0092233

Butler MA, King AA (2004) Phylogenetic comparative analysis: a modeling approach for adaptive evolution. Am Nat 164:683–695

Chiba S (2004) Ecological and morphological patterns in communities of land snails of the genus Mandarina from the Bonin Islands. J Evol Biol 17:131–143. doi:10.1046/j.1420-9101.2004.00639.x

Clabaut C, Bunje PME, Salzburger W, Meyer A (2007) Geometric morphometric analyses provide evidence for the adaptive character of the Tanganyikan cichlid fish radiations. Evolution 61:560–578. doi:10.1111/j.1558-5646.2007.00045.x

Cody ML (1974) Competition and the structure of bird communities, pp 1–318

Cody ML, Diamond JM (1975) Ecology and evolution of communities, pp 1–545

Cody ML, Mooney HA (1978) Convergence versus nonconvergence in Mediterranean-climate ecosystems. Annu Rev Ecol Syst 9:265–321. doi:10.1146/annurev.es.09.110178.001405

Collar DC, Schulte JA II, Losos JB (2011) Evolution of extreme body size disparity in monitor lizards (Varanus). Evolution 65:2664–2680. doi:10.1111/j.1558-5646.2011.01335.x

Conway Morris S (2003) Life's solution: inevitable humans in a lonely universe. pp 1–464

Cooper WJ, Westneat MW (2009) Form and function of damselfish skulls: rapid and repeated evolution into a limited number of trophic niches. BMC Evol Biol 9:24. doi:10.1186/1471-2148-9-24

De Busschere C, Baert L, Van Belleghem SM, Dekoninck W, Hendrickx F (2012) Parallel phenotypic evolution in a wolf spider radiation on Galápagos. Biol J Linn Soc 106:123–136. doi:10.1111/j.1095-8312.2011.01848.x

Eastman JM, Alfaro ME, Joyce P, Hipp AL, Harmon LJ (2011) A novel comparative method for identifying shifts in the rate of character evolution on trees. Evolution 65(12):3578–3589

Ellingson RA (2013) Convergent evolution of ecomorphological adaptations in geographically isolated Bay gobies (Teleostei: Gobionellidae) of the temperate North Pacific. Mol Phylogenet Evol. doi:10.1016/j.ympev.2013.10.009

Emerson BC, Gillespie RG (2008) Phylogenetic analysis of community assembly and structure over space and time. Trends Ecol Evol 23:619–630. doi:10.1016/j.tree.2008.07.005

Erwin DH (2006) Evolutionary contingency. Curr Biol 16:825–826

Felsenstein J (1988) Phylogenies and quantitative characters. Annu Rev Ecol Syst 19:445–471. doi:10.1146/annurev.es.19.110188.002305

Fox BJ (1987) Species assembly and the evolution of community structure. Evol Ecol 1:201–213. doi:10.1007/BF02067551

Frédérich B, Sorenson L, Santini F, Slater GJ, Alfaro ME (2013) Iterative ecological radiation and convergence during the evolutionary history of damselfishes (Pomacentridae). Am Nat 181:94–113. doi:10.1086/668599

Garland T, Dickerman AW, Janis CM, Jones JA (1993) Phylogenetic analysis of covariance by computer simulation. Syst Biol 42:265–292. doi:10.1093/sysbio/42.3.265

Gatz AJJ (1979) Community organization in fishes as indicated by morphological features. Ecology 60:711–718

Gavrilets S, Vose A (2005) Dynamic patterns of adaptive radiation. Proc Natl Acad Sci 102:18040–18045. doi:10.1073/pnas.0506330102

Gillespie R (2004) Community assembly through adaptive radiation in Hawaiian spiders. Science 303:356–359. doi:10.1126/science.1091875

Gillespie RG (2005) The ecology and evolution of Hawaiian spider communities. Am Sci 93:122–131

Givnish TJ (1997) Adaptive radiation and molecular systematics: aims and conceptual issues. In: Molecular evolution and adaptive radiation, Cambridge University Press, New York

Givnish TJ (1999) Adaptive radiation, dispersal, and diversification of the Hawaiian lobeliads. The Biology of Biodiversity. Springer, Tokyo

Givnish TJ, Millam KC, Mast AR, Paterson TB, Theim TJ, Hipp AL, Henss JM, Smith JF, Wood KR, Sytsma KJ (2009) Origin, adaptive radiation and diversification of the Hawaiian lobeliads (Asterales: Campanulaceae). Proc Roy Soc B 276:407–416. doi:10.1098/rspb.2008.1204

Glor RE (2010) Phylogenetic insights on adaptive radiation. Annu Rev Ecol Evol Syst 41:251–270. doi:10.1146/annurev.ecolsys.39.110707.173447

Glor RE, Kolbe JJ, Powell R, Larson A, Losos JB (2003) Phylogenetic analysis of ecological and morphological diversification in Hispaniolan trunk-ground anoles (Anolis cybotes group). Evolution 57:2383–2397

Gould SJ (1989) Wonderful life: the Burgess Shale and the nature of history, pp 1–347

Gould SJ (2002) The structure of evolutionary theory, pp 1–1464

Gould SJ (2003) Contingency. In: Palaeobiology II. Blackwell Publishing Ltd, New Jersey

Hansen TF (1997) Stabilizing selection and the comparative analysis of adaptation. Evolution 51:1341–1351

Hansen TF (2012) Adaptive landscapes and macroevolutionary dynamics. In: The adaptive landscape in evolutionary biology, Oxford University Press, Oxford

Hansen TF, Pienaar J, Orzack SH (2008) A comparative method for studying adaptation to a randomly evolving environment. Evolution 62:1965–1977. doi:10.1111/j.1558-5646.2008.00412.x

Harmon LJ, Kolbe JJ, Cheverud JM, Losos JB (2005) Convergence and the multidimensional niche. Evolution 59:409–421

Harmon LJ, Losos JB, Davies TJ, Gillespie RG, Gittleman JL, Jennings WB, Kozak KH, McPeek MA, Moreno-Roark F, Near TJ, Purvis A, Ricklefs RE, Schluter D, Schulte JA II, Seehausen O, Sidlauskas BL, Torres-Carvajal O, Weir JT, Mooers AØ (2010) Early bursts of body size and shape evolution are rare in comparative data. Evolution 64:2385–2396. doi:10.1111/j.1558-5646.2010.01025.x

Heath TA, Holder MT, Huelsenbeck JP (2012) A Dirichlet process prior for estimating lineage-specific substitution rates. Mol Biol Evol 29(3):939–955

Hekstra DR, Leibler S (2012) Contingency and statistical laws in replicate microbial closed ecosystems. Cell 149:1164–1173. doi:10.1016/j.cell.2012.03.040

Huelsenbeck JP, Nielsen R, Bollback JP (2003) Stochastic mapping of morphological characters. Syst Biol 52:131–158. doi:10.1080/10635150390192780

Ingram T, Kai Y (2014) The geography of morphological convergence in the radiations of Pacific *Sebastes* rockfishes

Ingram T, Mahler DL (2013) SURFACE: detecting convergent evolution from comparative data by fitting Ornstein-Uhlenbeck models with stepwise Akaike Information Criterion. Methods Ecol Evol 4:416–425. doi:10.1111/2041-210X.12034

Inkpen R, Turner D (2012) The topography of historical contingency. J Philos Hist 6:1–19. doi:10.1163/187226312X625573

Johnson MA, Revell LJ, Losos JB (2009) Behavioral convergence and adaptive radiation: effects of habitat use on territorial behavior in *Anolis* lizards. Evolution 64:1151–1159. doi:10.1111/j.1558-5646.2009.00881.x

Karr JR, James FC (1975) Eco-morphological configurations and convergent evolution in species and communities. In: Ecology and evolution of communities. The Belknap Press of Harvard University Press, Cambridge, MA

Kassen R (2009) Toward a general theory of adaptive radiation: insights from microbial experimental evolution. Ann N Y Acad Sci 1168:3–22. doi:10.1111/j.1749-6632.2009.04574.x

Keast A (1972) Ecological opportunities and dominant families, as illustrated by the Neotropical Tyrannidae (Aves). Evol Biol 5:229–277

Kelt DA, Brown JH, Heske EJ, Marquet PA, Morton SR, Reid JRW, Rogovin KA, Shenbrot G (1996) Community structure of desert small mammals: comparisons across four continents. Ecology 77:746–761

Kocher TD, Conroy JA, McKaye KR, Stauffer JR (1993) Similar morphologies of cichlid fish in Lakes Tanganyika and Malawi are due to convergence. Mol Phylogenet Evol 2:158–165

Kozak KH, Mendyk RW, Wiens JJ (2009) Can parallel diversification occur in sympatry? Repeated patterns of body-size evolution in coexisting clades of North American salamanders. Evolution 63:1769–1784. doi:10.1111/j.1558-5646.2009.00680.x

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

Lapiedra O, Sol D, Carranza S, Beaulieu JM (2013) Behavioural changes and the adaptive diversification of pigeons and doves. Proc Roy Soc B 280:20122893

Lawton JH (1984) Non-competitive populations, non-convergent communities, and vacant niches: The herbivores of bracken. In: Strong DRJ, Simberloff D, Abele LG, Thistle AB (eds) Ecological Communities: Conceptual Issues and the Evidence. Princeton University Press, Princeton, pp 67–100

Lenski RE, Travisano M (1994) Dynamics of adaptation and diversification: A 10,000-generation experiment with bacterial populations. Proc Natl Acad Sci USA 91:6808–6814

Losos JB (1996) Ecological and evolutionary determinants of the species-area relation in Caribbean anoline lizards. Philos Trans Roy Soc Lond B 351:847–854

Losos JB (2011) Convergence, adaptation and constraint. Evolution 65:1827–1840. doi:10.1111/j.1558-5646.2011.01289.x

Losos JB, Jackman TR, Larson A, de Queiroz K, Rodríguez-Schettino L (1998) Contingency and determinism in replicated adaptive radiations of island lizards. Science 279:2115–2118

Losos JB, Mahler DL (2010) Adaptive radiation: the interaction of ecological opportunity, adaptation, and speciation. In: Evolution since Darwin: The first 150 Years, vol 150. Sinauer Associates, Sunderland, MA

MacLean RC (2005) Adaptive radiation in microbial microcosms. J Evol Biol 18:1376–1386. doi:10.1111/j.1420-9101.2005.00931.x

Madsen O, Scally M, Douady CJ, Kao DJ, DeBry RW, Adkins R, Amrine HM, Stanhope MJ, de Jong WW, Springer MS (2001) Parallel adaptive radiations in two major clades of placental mammals. Nature 409:610–614. doi:10.1038/35054544

Mahler DL, Ingram T, Revell LJ, Losos JB (2013) Exceptional convergence on the macroevolutionary landscape in island lizard radiations. Science 341:292–295. doi:10.1126/science.1232392

Manceau M, Domingues VS, Linnen CR, Rosenblum EB, Hoekstra HE (2010) Convergence in pigmentation at multiple levels: mutations, genes and function. Philos Trans Roy Soc B: Biol Sci 365:2439–2450. doi:10.1098/rstb.2010.0104

Martins EP (1994) Estimating the rate of phenotypic evolution from comparative data. Am Nat 144:193–209

Melville J, Harmon LJ, Losos JB (2006) Intercontinental community convergence of ecology and morphology in desert lizards. Proc Roy Soc B 273:557–563. doi:10.1098/rspb.2005.3328

Miles DB, Ricklefs RE, Travis J (1987) Concordance of ecomorphological relationships in three assemblages of passerine birds. Am Nat 129:347–364

Moen DS, Wiens JJ (2009) Phylogenetic evidence for competitively driven divergence: body-size evolution in Caribbean treefrogs (Hylidae: *Osteopilus*). Evolution 63:195–214. doi:10.1111/j.1558-5646.2008.00538.x

Montaña CG, Winemiller KO (2013) Evolutionary convergence in Neotropical cichlids and Nearctic centrarchids: evidence from morphology, diet, and stable isotope analysis. Biol J Linn Soc 109:146–164

Muschick M, Indermaur A, Salzburger W (2012) Convergent evolution within an adaptive radiation of cichlid fishes. Curr Biol 22:2362–2368. doi:10.1016/j.cub.2012.10.048

Orians GH, Paine RT (1983) Convergent evolution at the community level. In: Coevolution, Sinauer Associates, Inc., Sunderland, MA

Orians GH, Solbrig OT (1977a) Convergent evolution in warm deserts. Dowden, Hutchinson & Ross, Inc., Stroudsburg, PA

Orians GH, Solbrig OT (1977b) Degree of convergence of ecosystem characteristics. In: Convergent evolution in warm deserts. Dowden, Hutchinson & Ross, Inc., Stroudsburg, PA

Patterson TB, Givnish TJ (2003) Geographic cohesion, chromosomal evolution, parallel adaptive radiations, and consequent floral adaptations in *Calochortus* (Calochortaceae): evidence from a cpDNA phylogeny. New Phytol 161:253–264. doi:10.1046/j.1469-8137.2003.00951.x

Pearce T (2012) Convergence and parallelism in evolution: a neo-Gouldian account. Br J Philos Sci 63:429–448. doi:10.1093/bjps/axr046

Pianka ER (1974) Evolutionary ecology, pp 1–397

Pie MR, Weitz JS (2005) A null model of morphospace occupation. Am Nat 166:E1–E13. doi:10.1086/430727

Powell R (2009) Contingency and convergence in macroevolution: a reply to John Beatty. J Philos 106:390–403

Powell R (2012) Convergent evolution and the limits of natural selection. Eur J Philos Sci 2:355–373. doi:10.1007/s13194-012-0047-9

Price T (1997) Correlated evolution and independent contrasts. Philos Trans Roy Soc Lond B 352:519–529

Price T, Lovette IJ, Bermingham E, Gibbs HL, Richman AD (2000) The imprint of history on communities of North American and Asian warblers. Am Nat 156:354–367. doi:10.1086/303397

Rainey PB, Travisano M (1998) Adaptive radiation in a heterogeneous environment. Nature 394:69–72

Revell LJ (2009) Size-correction and principal components for interspecific comparative studies. Evolution 63:3258–3268. doi:10.1111/j.1558-5646.2009.00804.x

Revell LJ (2012) phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol Evol 3(2):217–223

Revell LJ, Mahler DL, Peres-Neto PR, Redelings BD (2012) A new phylogenetic method for identifying exceptional phenotypic diversification. Evolution 66:135–146. doi:10.5061/dryad.vj310

Ricklefs RE, Miles DB (1994) Ecological and evolutionary inferences from morphology: an ecological perspective. University of Chicago Press, Chicago

Ricklefs RE, Schluter D (1993) Species diversity in ecological communities: historical and geographical perspectives, pp 1–414

Ricklefs RE, Travis J (1980) A morphological approach to the study of avian community organization. Auk 97:321–338

Rüber L, Verheyen E, Meyer A (1999) Replicated evolution of trophic specializations in an endemic cichlid fish lineage from Lake Tanganyika. Proc Natl Acad Sci 96:10230–10235

Ruedi M, Mayer F (2001) Molecular systematics of bats of the genus *Myotis* (Vespertilionidae) suggests deterministic ecomorphological convergences. Mol Phylogenet Evol 21:436–448. doi:10.1006/mpev.2001.1017

Sallan LC, Friedman M (2012) Heads or tails: staged diversification in vertebrate evolutionary radiations. Proc Roy Soc B: Biol Sci 279:2025–2032. doi:10.1098/rspb.2011.2454

Samuels CL, Drake JA (1997) Divergent perspectives on community convergence. Trends Ecol Evol 12:427–432

Saxer G, Doebeli M, Travisano M (2010) The repeatability of adaptive radiation during long-term experimental evolution of *Escherichia coli* in a multiple nutrient environment. PLoS ONE 5:e14184. doi:10.1371/journal.pone.0014184

Scheffer M, van Nes EH (2006) Self-organized similarity, the evolutionary emergence of groups of similar species. Proc Natl Acad Sci 103:6230–6235. doi:10.1073/pnas.0508024103

Schluter D (1986) Tests for similarity and convergence of finch communities. Ecology 67:1073–1085

Schluter D (1990) Species-for-species matching. Am Nat 136:560–568. doi:10.1086/285135

Schluter D (2000) The ecology of adaptive radiation, pp 1–288

Schluter D, McPhail JD (1993) Character displacement and replicate adaptive radiation. Trends Ecol Evol 8:197–200

Segar ST, Pereira RAS, Compton SG, Cook JM (2013) Convergent structure of multitrophic communities over three continents. Ecol Lett. doi:10.1111/ele.12183

Simpson GG (1944) Tempo and mode in evolution, pp 1–237

Simpson GG (1950) Evolutionary determinism and the fossil record. Sci Mon 71:262–267

Springer MS, Kirsch JAW, Chase JA (1997) The chronicle of marsupial evolution. In: Molecular evolution and adaptive radiation. Cambridge University Press, Cambridge

Stayton CT (2008) Is convergence surprising? An examination of the frequency of convergence in simulated datasets. J Theor Biol 252:1–14. doi:10.1016/j.jtbi.2008.01.008

Stiassny MLJ, Meyer A (1999) Cichlids of the rift lakes. Sci Am 280:64–69. doi:10.1038/scientificamerican0299-64

Streelman JT, Danley PD (2003) The stages of vertebrate evolutionary radiation. Trends Ecol Evol 18:126–131

Strong DRJ, Simberloff D, Abele LG, Thistle AB (1984) Ecological communities: conceptual issues and the evidence, pp 1–613

Svensson EI, Calsbeek R (2012) The past, the present, and the future of the adaptive landscape. In: Svensson EI, Calsbeek R (eds) The adaptive landscape in evolutionary biology. Oxford University Press, Oxford, pp 299–308

Thomas GH, Freckleton RP (2012) MOTMOT: models of trait macroevolution on trees. Methods Ecol Evol 3:145–151. doi:10.1111/j.2041-210X.2011.00132.x

Tyerman J, Havard N, Saxer G, Travisano M, Doebeli M (2005) Unparallel diversification in bacterial microcosms. Proc Roy Soc B 272:1393–1398. doi:10.1098/rspb.2005.3068

Uhlenbeck GE, Ornstein LS (1930) On the theory of the Brownian motion. Phys Rev 36:823–841

Venditti C, Meade A, Pagel M (2011) Multiple routes to mammalian diversity. Nature 479:393–396. doi:10.1038/nature10516

Wagenaar DA, Adami C (2004) Influence of chance, history, and adaptation on digital evolution. Artif Life 10:181–190. doi:10.1162/106454604773563603

Wiens JA (1991) Ecomorphological comparisons of the shrub-desert avifaunas of Australia and North America. Oikos 60:55–63

Wiens JJ, Brandley MC, Reeder TW (2006) Why does a trait evolve multiple times within a clade? Repeated evolution of snakelike body form in squamate reptiles. Evolution 60:123–141

Winchester JM, Boyer DM, St Clair EM, Gosselin-Ildari AD, Cooke SB, Ledogar JA (2014) Dental topography of platyrrhines and prosimians: convergence and contrasts. Am J Phys Anthropol 153:29–44. doi:10.1002/ajpa.22398

Yedid G, Bell G (2002) Macroevolution simulated with autonomously replicating computer programs. Nature 420:810–812. doi:10.1038/nature01151

Yedid G, Ofria CA, Lenski RE (2008) Historical and contingent factors affect re-evolution of a complex feature lost during mass extinction in communities of digital organisms. J Evol Biol 21:1335–1357. doi:10.1111/j.1420-9101.2008.01564.x

Young KA, Snoeks J, Seehausen O (2009) Morphological diversity and the roles of contingency, chance and determinism in African cichlid radiations. PLoS ONE 4:e4740. doi:10.1371/journal.pone.0004740

Chapter 19
Metrics and Models of Community Phylogenetics

**William D. Pearse, Andy Purvis, Jeannine Cavender-Bares
and Matthew R. Helmus**

Abstract Community phylogenetics combines ideas from community ecology and evolutionary biology, using species phylogeny to explore the processes underlying ecological community assembly. Here, we describe the development of the field's comparative methods and their roots in conservation biology, biodiversity quantification, and macroevolution. Next, we review the multitude of community phylogenetic structure metrics and place each into one of four classes: *shape*, *evenness*, *dispersion*, and *dissimilarity*. Shape metrics examine the structure of an assemblage phylogeny, while evenness metrics incorporate species abundances. Dispersion metrics examine assemblages given a phylogeny of species that could occupy those assemblages (the source pool), while dissimilarity metrics compare phylogenetic structure between assemblages. We then examine how metrics perform in simulated communities that vary in their phylogenetic structure. We provide an example of model-based approaches and argue that they are a promising area of future research in community phylogenetics. Code to reproduce all these analyses is available in the Online Practical Material (http://www.mpcm-evolution.com). We conclude by discussing future research directions for the field as a whole.

W. D. Pearse (✉) · J. Cavender-Bares
Department Ecology, Evolution, and Behavior, University of Minnesota,
1987 Upper Buford Circle, Saint Paul, MN 55108, USA
e-mail: will.pearse@gmail.com

J. Cavender-Bares
e-mail: cavender@umn.edu

A. Purvis
Department of Life Sciences, Natural History Museum,
Cromwell Road, London SW7 5BD, UK
e-mail: andy.purvis@nhm.ac.uk

M. R. Helmus
Amsterdam Global Change Institute, Department of Animal Ecology,
Vrije Universiteit, 1081 HV Amsterdam, The Netherlands
e-mail: mrhelmus@gmail.com

19.1 Overview

Community phylogenetics seeks to explore the ecological and evolutionary factors that underlie the assembly of communities and how species interactions influence evolutionary and ecosystem processes. The field represents a (re-)integration of community ecology and evolution, in the hope that historical species interactions and environmental conditions reflected in phylogeny can inform us about present-day ecology. However, rapid advances in computational tools, phylogenetic inference methods, DNA databases, and metrics mean the scope of community phylogenetics is constantly expanding and developing.

This chapter should provide the reader with an entry point to begin critically conducting their own community phylogenetic analysis. To this end, the Online Practical Material (http://www.mpcm-evolution.com) contains annotated R (R Core Team 2014) code with which the reader can repeat all the analyses and simulations presented in this chapter. We begin by describing the development of community phylogenetics and follow by outlining a framework to understand community phylogenetic metrics. We then examine the performance of several metrics in a simulated data set and give a brief introduction to the field of community phylogenetic modelling. We conclude the chapter by discussing caveats and future directions for the field.

19.2 Historical Overview of the Metrics of Community Phylogenetics

It was Darwin 1859 who first hypothesised a relationship between species' taxonomic proximity and competitive interactions, arguing that congeners use the same resources and so competition should be strongest among them. While Darwin was interested in how this increasing competition would affect natural selection, later scientists (Jaccard 1901; Elton 1946) would ask how the number of congeners present in a community reflected the biogeographic and ecological processes structuring it. Despite controversies over the sensitivity of such approaches to species richness (Järvinen 1982), the idea that ecological processes could be detected in the evolutionary relationships among species in ecological communities took hold.

Conservation biologists were quick to recognise the utility of phylogeny as a way to quantify species uniqueness and thus aid conservation prioritisation. Vane-Wright et al. (1991) first argued to prioritise more basal evolutionary lineages (acknowledged by May (1990) who published first), and Altschul and Lipman (1990) suggested incorporating time-calibrated phylogenies and Felsenstein's comparative method (1985). Soon after, Faith (1992) coined the phylogenetic diversity (PD) metric as the summed phylogenetic branch length connecting all species in a set to rank areas for preservation. Later metrics partitioned the

phylogenetic diversity of clades among their species to facilitate species-based conservation of phylogenetic diversity (Pavoine et al. 2005; Redding and Mooers 2006; Isaac et al. 2007).

In parallel, the taxonomy-based metrics developed in conservation biology (May 1990; Vane-Wright et al. 1991) were adapted to understand community assembly in degraded ecosystems. Warwick and Clarke (1995) counted the mean number of taxonomic ranks separating community members to derive the Δ family of metrics (some members were independently derived by Izsáki and Papp 1995), which were later extended to estimate taxonomic similarity among communities (Izsáki and Price 2001). Although not the first modern study of ecological taxonomic structure (cf., e.g. Douglas and Matthews 1992), Warwick and Clarke (1995) provided the first clear example of how habitat filtering can change the taxonomic (and so phylogenetic) composition of ecological communities.

These antecedents provided the basis for papers (Webb 2000; Webb et al. 2002) that developed a framework and set of hypotheses for the use of phylogenetics in mainstream ecology and mark the beginning of modern community phylogenetics. Webb (2000) developed the Net Relatedness Index (NRI) and the Nearest Taxon Index (NTI) to measure the phylogenetic structure of a tropical forest plot. NRI and NTI examine whether the relatedness of species to one another in a community differs from what would be expected under random assembly from a list of potential species (the source pool). Most community phylogenetic studies assume close relatives are ecologically similar (niche conservatism; reviewed in Wiens et al. 2010). Under this assumption, communities whose species are more closely related than under random assembly (underdispersed, or clustered, communities) reflect habitat-filtered assembly, while communities of unexpectedly distantly related species (overdispersed communities) indicate the influence of competitive exclusion.

The assumption of niche conservatism has subsequently been scrutinised, and inferring ecological process purely on the basis of phylogenetic pattern is now treated with scepticism. For example, in one of the early empirical tests of the Webb et al. (2002) framework, Cavender-Bares et al. (2004) demonstrated that niche convergence (rather than conservatism) among distantly related oak lineages caused overdispersion in hyper-diverse oak forest communities. Subsequently, Valiente-Banuet and Verdú (2007) found that facilitation among distantly related species could lead to overdispersion, and Mayfield and Levine (2010) argued that competition may lead to phylogenetic clustering. The ecological and evolutionary mechanisms that produce phylogenetic community structure vary and depend on where in the tree of life one is looking (phylogenetic and biogeographic scale) and the modes of trait evolution at work (Cavender-Bares et al. 2006). Kraft et al. (2007) demonstrated that when known ecological and evolutionary processes are simulated, the anticipated community phylogenetic patterns are reliably recovered, but Kembel (2009) has shown that dispersal can mask such patterns. Development of model-based methods (see Sect. 19.5) offers hope of explicitly testing mechanistic hypotheses about how evolutionary and ecological processes interact, reconciling many objections about inferring process from phylogenetic pattern.
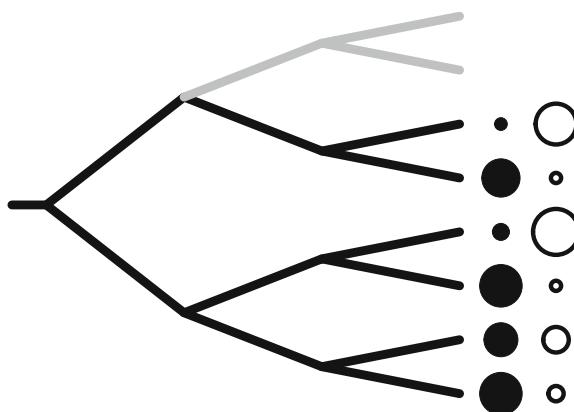
Fig. 19.1 Overview of phylogenetic *shape*, *evenness*, *dispersion*, and *dissimilarity* metrics, as described in Sect. 19.3. *Shape* metrics measure only the observed assemblage phylogeny—the parts of the phylogeny in *black*. *Evenness* metrics measure how evenly species' abundances are distributed across the assemblage phylogeny; the abundances of species in two communities are represented by the size of filled and open *circles* on the figure. *Dispersion* metrics examine whether the observed members of an assemblage are a random subset of the species pool (*grey* and *black* parts of the phylogeny). *Dissimilarity* metrics quantify phylogenetic similarity between observed assemblages. The two assemblages in this figure contain the same species, and so their phylogenetic dissimilarity is null unless abundances are taken into account

19.3 A Systematic Classification of Community Phylogenetic Metrics

While many have reviewed issues in community phylogenetics (e.g. Emerson and Gillespie 2008; Graham and Fine 2008; Cavender-Bares et al. 2009; Vamosi et al. 2009; Mouquet et al. 2012; Swenson 2013), our focus here is specifically on methodology. Pavoine and Bonsall (2011) are of note in that they emphasise analogues between phylogenetic and functional trait diversity and define six major classes of diversity metric. Three of these classes (all shape measures in our classification) are of particular interest here: *multivariate richness* (the sum of a phylogeny's branches; essentially *PD*), *regularity* (the balance of a tree; see Sect. 19.3.1), and *divergence* (the mean distance among species). Vellend et al. (2011) chose a very different scheme, classifying phylogenetic diversity metrics as '*type I*' or '*type II*' depending on whether they begin by measuring the phylogenetic distinctiveness of species (*I*) or examine subsets of a regional phylogeny (*II*).

We propose four classes of community phylogenetic structure with names chosen to reflect existing community ecological literature: shape, evenness, dispersion, and dissimilarity. A graphical overview of these measures is given in Fig. 19.1, and more than 40 metrics are placed into the scheme in the online supplementary materials. *Shape* metrics describe an assemblage phylogeny's topology, branch lengths, size, how closely related its species are, and many

predate community phylogenetics. *Evenness* metrics reflect how species' abundances are distributed throughout a phylogeny, and many are extensions of existing metrics of species diversity. *Dispersion* metrics ask whether an assemblage phylogeny differs from what would be expected, under a given null model, from a source pool phylogeny of potential and actual members of that assemblage. Finally, *dissimilarity* metrics quantify differences in the phylogenetic composition of species occupancy and abundance between assemblages.

19.3.1 Shape

Shape metrics assess the structure of a phylogeny alone and can be calculated with only a list of species and their phylogeny (reviewed in Mooers and Heard 1997). Many predate community phylogenetics itself and were intended for use in macroevolutionary studies. One of the more well known is Colless' Index (I_C), which measures phylogenetic balance as the extent to which nodes in a phylogeny define subgroups of equal size (Colless 1982). An unbalanced assemblage phylogeny indicates that particular clades dominate that assemblage, perhaps because they display key traits that adapt them to that environment. The γ statistic (Pybus and Harvey 2000) was originally intended to detect decreases in the rate of diversification through time; in a community phylogenetic context, this is consistent with an assemblage containing species that are relatively unrelated to one another. Phylogenetic species richness (PSV; Helmus et al. 2007) measures whether the distribution of species across the phylogeny differs from expectation under a Brownian null model and is analogous to the mean phylogenetic distance (MPD) among species on a phylogeny. We caution that an assemblage phylogeny is affected by processes operating outside the assemblage (see Heard and Cox 2007); shape measures sensitive to symmetry at different phylogenetic depths (see Agapow and Purvis 2002) may be useful tools when exploring these issues.

19.3.2 Evenness

Measures of evenness ask whether species abundances are biased towards any particular clade(s) throughout the phylogeny. Many are extensions of existing measures of ecological diversity or shape measures; for instance, the Imbalance in Abundance of higher Clades (IAC; Cadotte et al. 2010) metric is essentially an abundance-weighted form of I_C. Classical measures of the phylogenetic signal of species' traits (e.g. Pagel's λ; 1999) are evenness metrics when calculated using species' abundances, although in most cases statistical transformation of abundances (e.g. taking their logarithm) is advised. Often shape and evenness metrics are calculated for individual sites (α shape/evenness) and across a landscape (γ) to measure β shape or evenness (see Graham and Fine 2008). We intend to use the

term *evenness* analogously to its use in other fields of ecology (see Magurran 2004; Pavoine and Bonsall 2011), but the reader should note that Kraft et al. (2007) and others use the term 'phylogenetic evenness' to indicate communities that contain more distantly related species than expected from null models. We suggest the use of the term 'phylogenetic overdispersion' for this case in reference to the statistical definition of overdispersion (see Sect. 19.3.3 below).

19.3.3 Dispersion

Metrics of phylogenetic dispersion describe whether an observed assemblage is a phylogenetically biased subset of the species that could coexist in that assemblage (the source pool). Bias can reflect community assembly or survival of an extinction episode, and most metrics focus on whether individuals or species are more or less related to one another (under- or overdispersed, respectively) than under a null expectation. They differ from shape and evenness measures, upon which they are often based, in that they require a null expectation; their value is contingent not just upon the observed assemblage but also a null expectation derived from random assembly of same-sized assemblages from a more inclusive source pool. NRI and NTI are the best known: the Net Relatedness Index (NRI) compares the phylogenetic distance among all members of a community, while the Nearest Taxon Index (NTI) examines only distances among nearest relatives. The first definition of NRI and NTI (Webb 2000) counted nodal distance between species, and the second (Webb et al. 2002) used phylogenetic branch lengths. Kembel (2009) defined standard effect sizes of MPD and the mean nearest taxon distance (SES_{MPD} and SES_{MNTD}), which are the negations of NRI and NTI, respectively. Pearse et al. (2013) showed that the randomisations that control for phylogenetic structure in NRI and NTI make the measures test statistics, and so their absolute values can be misleading. They found that D (Fritz and Purvis 2010), which is based upon independent contrasts (Felsenstein 1985) and a Brownian null distribution, can be more sensitive than NRI.

19.3.4 Dissimilarity

Measures of dissimilarity explicitly examine differences in assemblages' compositions, and many have analogues with classical ecological measures (e.g. *PhyloSor* and Sørensen's Index; Bryant et al. 2008). Unlike standard dissimilarity metrics, phylogenetic dissimilarity metrics differentiate among communities with no shared species. The metric phylogenetic community dissimilarity (PCD; Ives and Helmus 2010), for example, partitions dissimilarity into compositional (the proportion of shared species) and phylogenetic (the relatedness of unshared species) components, but the most widely used measure—especially by microbial

ecologists—is *UniFrac* (Lozupone and Knight 2005). *UniFrac* measures the amount of phylogenetic branch length unique to each community, essentially asking how much PD is unique to each community. These measures are distinct from measures of co-evolution (see Chap. 20) in comparing assemblages across a common phylogeny.

19.4 Quantitative Classification of Community Phylogenetic Metrics

Our classification of metrics into four groups is based on how the metrics are calculated and what the metrics attempt to measure. Here, we ask whether the members of our groupings give similar results in common data sets.

Given a particular number of species, n, in a species pool, there is a finite number of unique community compositions that we label the feasible set of community compositions for n species (Haegeman and Loreau 2008; Locey and White 2013). The feasible set can be calculated for any n (though it is often approximated for large n), and given a phylogeny, the distribution of the feasible values of any phylogenetic metric can be derived. Figure 19.2a provides the feasible distribution of PSV for a fully balanced phylogeny of 8 species with equal branch lengths (as in Fig. 19.1). Using this same tree, we simulated 6,000 communities, half structured by phylogenetic attraction where closely related species were more likely to be found together and half by repulsion (the converse). For an ultrametric phylogeny of n species with covariance matrix $\mathbf{V}$, we defined attraction as the Cholesky decomposition of $\mathbf{V}$ and repulsion as the decomposition of $\mathbf{V}^{-1}$, referring to either as the matrix $\mathbf{L}$ below. The probability of species s residing in a simulated community was a stochastic process as defined by

$$p_s = \frac{e^{c\mathbf{LR}}}{1 + e^{c\mathbf{LR}}} \tag{19.1}$$

where c was a scalar (fixed at 10) that determined the strength of attraction/repulsion and R an $n \times 1$ matrix of normally distributed random numbers centred at 0. The PSV distributions of these two simulated communities (Fig. 19.2b, c) differ markedly from the feasible set (Fig. 19.2a) and from each other, suggesting that metric distributions for empirical and feasible sets of communities can be compared to detect processes that cause phylogenetic attraction and repulsion. To group metrics calculated on these three data sets, we obtained the values of 27 metrics across the simulated communities and hierarchically clustered the metrics (*R* function *hclust*, complete linkage method) based on their standardised (centred to have a mean of zero and standard deviation of 1) Euclidian distances. Our methods were chosen to permit direct comparison with a similar study by Cadotte et al. (2010).
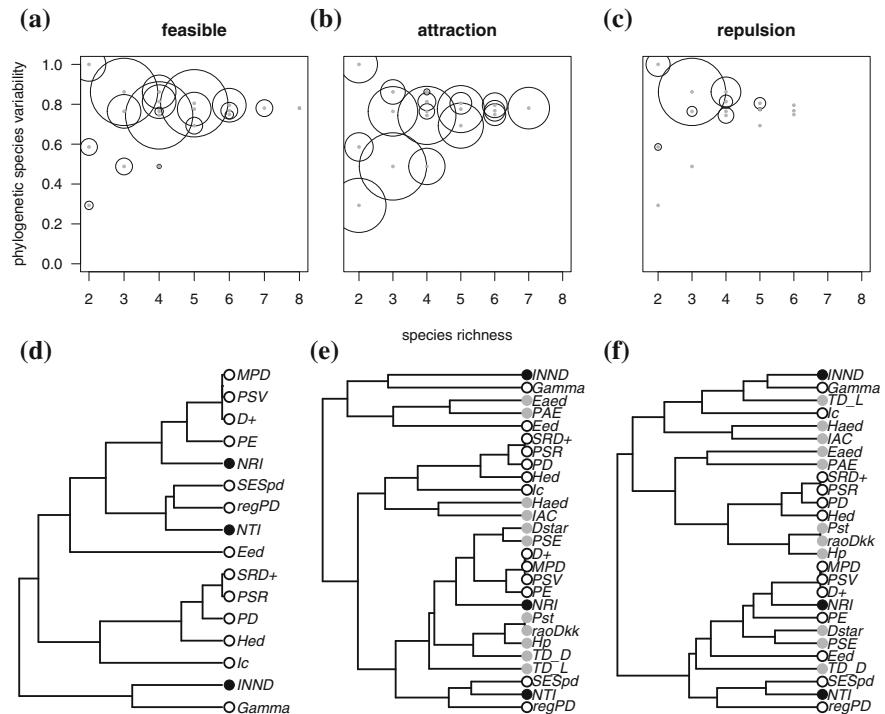
Fig. 19.2 Distributions and clustering dendrograms of community phylogenetic metrics calculated for feasible communities and communities simulated under models of phylogenetic attraction or repulsion. In (**a**), (**b**), and (**c**), the size of the circles (centres marked in *grey*) represent the numbers of unique species compositions that give each PSV value. **a** The feasible distribution of PSV for an eight species, balanced phylogeny. **b** The PSV distribution for communities simulated with attraction is generally lower than the feasible distribution and much lower than the communities simulated under repulsion (**c**). Below each distribution (**d**–**f**) are dendrograms based on a hierarchical clustering of the values of 27 community phylogenetic metrics calculated for each data set. The number of metrics differs among (**d**) and (**e**, **f**) because (**d**) uses the feasible set of species which is defined only for species presence/absence; in (**e**) and (**f**), we simulated abundances and thus incorporated evenness metrics. The *white*, *black*, and *grey* circles indicate shape, dispersion, and evenness metrics, respectively. See the Online Electronic Material for all metric names and abbreviations

The classification of Sect. 19.3 does not perfectly map onto groupings in Fig. 19.2d, e, f; the clustering of the metrics was inconsistent between the attraction and repulsion simulations. This suggests that a single quantitative classification of metrics is unlikely since the metric correlations depend on the underlying data set, but the systematic classification of Sect. 19.3 provides definitive categories for all the metrics.

19.5 Statistical Models of Community Phylogenetic Structure

While community phylogenetic metrics will continue to be developed, explicit statistical models are the next methodological frontier. Models explain phylogenetic structure across a number of assemblages simultaneously, maximising statistical power, and can incorporate phylogenetic, environmental, trait, and other information. While not models in the statistical sense of fitting probability distributions, the first model-based approach stemmed from randomisation methods developed to infer meta-community processes (Pillar and Duarte 2010). There are a number of related approaches (Leibold et al. 2010; Pavoine and Bonsall 2011; Peres-Neto et al. 2012), most involving the comparison of site-by-species matrices with matrices of environmental and species trait data.

Fitting a statistical model to community data allows for estimates of covariate effects and their errors, prediction of community composition, and model comparison using test statistics. Phylogenetic generalised linear mixed models (PGLMMs; Ives and Helmus 2011) were the first statistical models to be developed for community phylogenetics. Here, we illustrate the simplest PGLMM that predicts community composition on the basis of phylogeny alone. Fitting more complex models is useful and possible, but comes with a greater computational complexity and risk of fitting models too complex to be parameterised from the data at hand.

For n species distributed across m sites, the probability of any species being found at a site is logistically modelled as follows:

$$\mu_i = \text{logit}^{-1}(\alpha_{\text{spp}} + b_i + c_{\text{site}}) \tag{19.2}$$

where i indexes a particular spp at a particular *site*, α_{spp} is a categorical fixed effect that accounts for variation in species prevalence across communities, and b_i is a Gaussian distributed random effect with mean 0 that accounts for phylogeny. The covariance matrix of b_i is the Kronecker product of $\mathbf{I_m}$ and $\sigma^2_{\text{spp}}\mathbf{V_{spp}}$, where $\mathbf{I_m}$ is the $m \times m$ identity matrix and $\sigma^2_{\text{spp}}\mathbf{V_{spp}}$ an estimated scalar multiplied by the $n \times n$ phylogenetic covariance matrix. The resulting covariance matrix of b_i is block diagonal with $\sigma^2_{\text{spp}}\mathbf{V_{spp}}$ repeated as the blocks and zeroes elsewhere. Lastly, c_{site} is similar to b but with ones as the blocks in its covariance matrix. Including α_{spp} and c_{site} separates differences in species prevalence across communities or community size from phylogenetic effects.

We fit the PGLMM in Eq. 19.2 to the two simulated data sets from Sect. 19.4 (depicted in Fig. 19.2b and c). Note that the estimated scalar σ^2_{spp} gives the strength of the phylogenetic attraction, not repulsion, and so we expected σ^2_{spp} to only be significant for the communities simulated with attraction. Indeed, we identified significant phylogenetic pattern only in the attraction-simulated communities (attraction σ^2_{spp}: 0.87, 0.78–0.95 95 % CI, repulsion σ^2_{spp}: 0.0, 0.0–0.1 95 % CI). To test for repulsion, we altered the covariance matrix of b_i by replacing

$\sigma_{\text{spp}}\mathbf{V}_{\text{spp}}$ with $\sigma_{\text{spp}}\mathbf{V}_{\text{spp}}^{-1}$ (i.e. we replaced b_i with d_i from model III of Ives and Helmus 2011) and fit this new model to both data sets (note the similarity with Sect. 19.4 model). This second PGLMM only detected phylogenetic pattern in the communities simulated under the repulsion model (attraction σ_{spp}^2: 0.0, 0.0–0.10 95 % CI, repulsion σ_{spp}^2: 0.96, 0.87–1.06 95 % CI). The nature of this PGLMM and its performance detecting phylogenetic dispersion did not change, as many metrics did in Sect. 19.4. However, these models were calculated across only the first 50 of the communities simulated in Sect. 19.4 due to computational limitations, although new algorithms may overcome this (Ho and Ané 2014).

19.6 Future Developments

Studies often use trait data to justify investigators' assumption of niche conservatism and thus map phylogenetic pattern onto ecological process (which is contentious at best; Cavender-Bares et al. 2009). Yet if phylogeny is only a proxy for species traits, it is unclear why a phylogenetic 'middleman' (Swenson 2013) is needed when the trait data themselves are available. If we are to claim that a perfect phylogeny reflects species' niches better than trait data (Srivastava et al. 2012) or that phylogeny is a useful proxy for difficult to obtain functional trait data (Mace et al. 2003), then we must directly compare the explanatory power of traits and phylogeny. Yet phylogenetic signal in a trait does not mean phylogenetic and trait data are in perfect agreement; even if they were (not), measurement error may falsely indicate (dis)agreement. Recent developments, such as the *traitgram* approach (Cadotte et al. 2013), allow the explanatory power of phylogeny and traits to be partitioned and interactions between the two explored. An alternative is to contrast the evolution of species' traits with their present-day ecology; Cavender-Bares et al. (2006) found that traits critical to habitat filtering (such as plant height) were convergent, while traits associated with local competition (such as leaf habit) were conserved in oak trees. Silvertown et al. (2006; also see Ackerly et al. 2006) went a step further, categorising traits as α, β, or γ depending on their order of evolutionary divergence and relating these evolutionary dynamics to the likelihood of species coexisting in the present.

Community phylogenetics provides an excellent framework within which to examine the 'problem and promise of scale dependency' (Swenson et al. 2006), and spatial and taxonomic scaling continues to draw interest (e.g. Cavender-Bares et al. 2006; Kembel and Hubbell 2006). There is evidence of variation among clades in phylogenetic structure even within well-defined groups (e.g. Parra et al. 2010) and tentative evidence of links between clade age and phylogenetic dispersion (Pearse et al. 2013). Variation among clades is to be expected; under a Brownian model of trait evolution (which most metrics assume or, like PSV, are derived under; see also Peres-Neto et al. 2012), phylogeny is a poorer predictor of similarity for distantly related species. Advances in the modelling of species' trait evolution (reviewed in see also Cooper et al. 2010; Chap. 14) have provided us

with more sophisticated models of trait evolution, which should generate different expectations for ecological dissimilarity and so present-day phylogenetic structure. Indeed, the mode of speciation in a clade could affect its community phylogenetic structure today. To give a simplified example, species brought back into secondary contact may be unlikely to coexist due to shared environmental tolerance (or gene flow; see Fig. 5 in Cavender-Bares et al. 2009), while descendents of rapid adaptive radiations might be sufficiently dissimilar to coexist.

19.7 Conclusion

There will never be one perfect definition of an ecological assemblage, and so there will never be one perfect way of describing one. It is no surprise that some express misgivings about the incursion of phylogenetic structure into ecology; initial attempts to incorporate phylogenies into comparative analysis were met with criticism, and many feared that implicit assumptions of the approach were ignored (e.g. Westoby et al. 1995). Such initial scepticism is healthy—there is always a danger that a new framework will be applied simply because it can be, without any critical evaluation of its implications. The incorporation of phylogenetic structure into ecology is not without its pitfalls, but a little over a decade since Webb et al. (2002) outlined their research paradigm we have a remarkably mature suite of metrics and methods. Looking forward, ecologists and evolutionary biologists are moving beyond describing phylogenetic structure, and instead, testing detailed hypotheses about how that structure came to be. Species' evolutionary history was shaped by their ecology, and it seems natural to see what the shape of species' evolutionary past can reveal about their ecology today.

References

Ackerly DD, Schwilk DW, Webb CO (2006) Niche evolution and adaptive radiation: testing the order of trait divergence. Ecology 87:S50–S61

Agapow P-M, Purvis A (2002) Power of eight tree shape statistics to detect nonrandom diversification: a comparison by simulation of two models of cladogenesis. Syst Biol 51(6):866–872

Altschul SF, Lipman DJ (1990) Equal animals. Nature 348:493–494

Bryant JA et al (2008) Microbes on mountainsides: contrasting elevational patterns of bacterial and plant diversity. Proc Natl Acad Sci 105(S1):11505–11511

Cadotte M, Albert CH, Walker SC (2013) The ecology of differences: assessing community assembly with trait and evolutionary distances. Ecol Lett 16(10):1234–1244

Cadotte MW et al (2010) Phylogenetic diversity metrics for ecological communities: integrating species richness, abundance and evolutionary history. Ecol Lett 13(1):96–105

Cavender-Bares J, Keen A, Miles B (2006) Phylogenetic structure of Floridian plant communities depends on taxonomic and spatial scale. Ecology 87(7):S109–S122

Cavender-Bares J et al (2004) Phylogenetic overdispersion in Floridian oak communities. Am Nat 163(6):823–843

Cavender-Bares J et al (2009) The merging of community ecology and phylogenetic biology. Ecol Lett 12:693–715

Colless DH (1982) Review of phylogenetics: the theory and practice of phylogenetic systematics. Syst Zool 31(1):100–104

Cooper N, Jetz W, Freckleton RP (2010) Phylogenetic comparative approaches for studying niche conservatism. J Evol Biol 23(12):2529–2539

Darwin C (1859) On the origin of species. John Murray, London

Douglas ME, Matthews WJ (1992) Does morphology predict ecology? Hypothesis testing within a freshwater stream fish assemblage. Oikos 65(2):213–224

Elton C (1946) Competition and the structure of ecological communities. J Anim Ecol 15(1):54–68

Emerson BC, Gillespie RG (2008) Phylogenetic analysis of community assembly and structure over space and time. Trends Ecol Evol 23(11):619–630

Faith DP (1992) Conservation evaluation and phylogenetic diversity. Biol Conserv 61(1):1–10

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

Fritz SA, Purvis A (2010) Selectivity in mammalian extinction risk and threat types: a new measure of phylogenetic signal strength in binary traits. Conserv Biol 24(4):1042–1051

Graham CH, Fine PVA (2008) Phylogenetic beta diversity: linking ecological and evolutionary processes across space in time. Ecol Lett 11(12):1265–1277

Haegeman B, Loreau M (2008) Limitations of entropy maximization in ecology. Oikos 117:1700–1710

Heard SB, Cox GH (2007) The shapes of phylogenetic trees of clades, faunas, and local assemblages: exploring spatial pattern in differential diversification. Am Nat 169(5):E107–E118

Helmus MR et al (2007) Phylogenetic measures of biodiversity. Am Nat 169(3):E68–E83

Ho LST, Ane C (2014). A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. Syst Biol 63:397–408

Isaac NJB et al (2007) Mammals on the EDGE: conservation priorities based on threat and phylogeny. PLoS ONE 2(3):e296

Ives AR, Helmus MR (2010) Phylogenetic metrics of community similarity. Am Nat 176(5):E128–E142

Ives AR, Helmus MR (2011) Generalized linear mixed models for phylogenetic analyses of community structure. Ecol Monogr 81(3):511–525

Izsak C, Price ARG (2001) Measuring β-diversity using a taxonomic similarity index, and its relation to spatial scale. Mar Ecol Prog Ser 215:69–77

Izsáki J, Papp L (1995) Application of the quadratic entropy indices for diversity studies of drosophilid assemblages. Environ Ecol Stat 2(3):213–224

Jaccard P (1901) Etude comparative de la distribution florale dans une portion des Alpes et des Jura. Bull de la Soc Vaudoise des Sci Nat 37:547–579

Järvinen O (1982) Species-to-genus ratios in biogeography: a historical note. J Biogeogr 9(4):363–370

Kembel SW (2009) Disentangling niche and neutral influences on community assembly: assessing the performance of community phylogenetic structure tests. Ecol Lett 12(9):949–960

Kembel SW, Hubbell SP (2006) The phylogenetic structure of a neotropical forest tree community. Ecology 87(7):S86–S99

Kraft NJB et al (2007) Trait evolution, community assembly, and the phylogenetic structure of ecological communities. Am Nat 170(2):271–283

Leibold MA, Economo EP, Peres-Neto P (2010) Metacommunity phylogenetics: separating the roles of environmental filters and historical biogeography. Ecol Lett 13(10):1290–1299

Locey KJ, White EP (2013) How species richness and total abundance constrain the distribution of abundance. Ecol Lett 16(9):1177–1185

Lozupone C, Knight R (2005) UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol 71(12):8228–8235

Mace GM, Gittleman JL, Purvis A (2003) Preserving the tree of life. Science 300(5626):1707–1709

Magurran AE (2004). Measuring biological diversity. Oxford University Press, Oxford

May RM (1990) Taxonomy as destiny. Nature 347:129–130

Mayfield MM, Levine JM (2010) Opposing effects of competitive exclusion on the phylogenetic structure of communities. Ecol Lett 13(9):1085–1093

Mooers AØ, Heard SB (1997) Inferring evolutionary process from phylogenetic tree shape. Q Rev Biol 72(1):31–54

Mouquet N et al (2012) Ecophylogenetics: advances and perspectives. Biol Rev 87(4):769–785

Pagel M (1999) Inferring the historical patterns of biological evolution. Nature 401(6756):877–884

Parra JL, McGuire JA, Graham CH (2010) Incorporating clade identity in analyses of phylogenetic community structure: an example with hummingbirds. Am Nat 176(5):573–587

Pavoine S, Bonsall MB (2011) Measuring biodiversity to explain community assembly: a unified approach. Biol Rev 86(4):792–812

Pavoine S, Ollier S, Dufour A-B (2005) Is the originality of a species measurable? Ecol Lett 8(6):579–586

Pearse WD, Jones A, Purvis A (2013) Barro Colorado Island's phylogenetic assemblage structure across fine spatial scales and among clades of different ages. Ecology 94(12):2861–2872

Peres-Neto PR, Leibold MA, Dray S (2012) Assessing the effects of spatial contingency and environmental filtering on metacommunity phylogenetics. Ecology 93:S14–S30

Pillar VD, Duarte LS (2010) A framework for metacommunity analysis of phylogenetic structure. Ecol Lett 13(5):587–596

Pybus OG, Harvey PH (2000) Testing macro-evolutionary models using incomplete molecular phylogenies. Proc Roy Soc B Biol Sci 267(1459):2267–2272

R Core Team (2014) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Redding DW, Mooers AØ (2006) Incorporating evolutionary measures into conservation prioritization. Conserv Biol 20(6):1670–1678

Silvertown J et al (2006) Phylogeny and the hierarchical organization of plant diversity. Ecology 87(7):S39–S166

Srivastava DS et al (2012) Phylogenetic diversity and the functioning of ecosystems. Ecol Lett 15(7):637–648

Swenson NG (2013) The assembly of tropical tree communities—the advances and shortcomings of phylogenetic and functional trait analyses. Ecography 36(3):264–276

Swenson NG et al (2006) The problem and promise of scale dependency in community phylogenetics. Ecology 87(10):2418–2424

Valiente-Banuet A, Verdú M (2007) Facilitation can increase the phylogenetic diversity of plant communities. Ecol Lett 10(11):1029–1036

Vamosi S et al (2009) Emerging patterns in the comparative analysis of phylogenetic community structure. Mol Ecol 18(4):572–592

Vane-Wright RI, Humphries CJ, Williams PH (1991) What to protect?—systematics and the agony of choice. Biol Conserv 55(3):235–254

Vellend M et al. (2011) "Measuring phylogenetic biodiversity". In: Magurran AE, McGill BJ Biological Diversity, Oxford University Press, Oxford Chap. 14

Warwick RM, Clarke KR (1995) New 'biodiversity' measures reveal a decrease in taxonomic distinctness with increasing stress. Mar Ecol Prog Ser 129(1):301–305

Webb CO (2000) Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. Am Nat 156(2):145–155

Webb CO et al (2002) Phylogenies and community ecology. Annu Rev Ecol Syst 33(1):475–505

Westoby M, Leishman MR, Lord JM (1995) On misinterpreting the 'phylogenetic correction'. J Ecol 83(3):531–534

Wiens JJ et al (2010) Niche conservatism as an emerging principle in ecology and conservation biology. Ecol Lett 13(10):1310–1324

Chapter 20
Event-Based Cophylogenetic Comparative Analysis

Michael Charleston and Ran Libeskind-Hadas

Abstract Cophylogenetic analysis seeks to explain the relationships between mutually evolving pairs of species such as hosts and parasites. In the last two decades, increasingly sophisticated computational methods have been developed for performing cophylogenetic analyses. In particular, event-based reconstruction methods attempt to find the best supported reconstructions of pairs of related trees using a set of events including cospeciation, duplication, transfer, and loss. This chapter formulates the cophylogeny reconstruction problem, describes the algorithmic techniques that have been developed for this problem, and compares and contrasts the software packages that implement these methods.

20.1 Introduction

It has been famously said and restated many times that nothing in biology makes sense except in light of evolution (Dobzhansky 1973). We would add another axiom: *nothing evolves in isolation*. In every biological system, there are interactions among individuals, between strains and species, across genera and ecological roles. It is as much an unavoidable outcome of biological life as is evolution itself: All biological systems evolve; all biological systems include interactions; interactions influence fitness; and differences in fitness influence natural selection—evolution therefore cannot make sense except in light of *co*evolution.

M. Charleston
School of Information Technologies, University of Sydney, Sydney, Australia
e-mail: michael.charleston@sydney.edu.au

R. Libeskind-Hadas (✉)
Department of Computer Science, Harvey Mudd College, Claremount, USA
e-mail: ran@cs.hmc.edu

Coevolution can mean many things: from the very "small-scale" changes in the *env* gene giving rise to conformational changes of the HIV virus capsid in response to changing drug therapies in an AIDS patient (Chun et al. 1999), to phenotypic changes of large ectoparasites, e.g., lice in response to diversification of their host species (Hafner and Nadler 1988).

The term "coevolution" is sometimes assumed to be strictly reciprocal: Two organisms exert selective pressure on one another and both evolve in response (Thompson 1994). This may arise, for example, in host–parasite relationships, mutualisms, and competitive relationships. More generally, coevolution is a statement about non-independence of pairs of species: At least one species—or, more generally, *taxonomic unit* or *taxon*—is evolving non-independently of another.

In this chapter, we focus on the quite common asymmetric relationship between groups of ecologically linked taxa such as that which exists between pathogens or parasites and their hosts. The relationship is asymmetric because in cases such as these, which represent a strong majority of cases in which coevolutionary analysis is applied, one organism is evolving much more slowly than the other. It should be noted that some exceptions to this rule are known to exist (Cuthill and Charleston 2012).

For example, in a now very famous exemplar of coevolution at the species level, Hafner and Nadler's "gopher-louse" study (Hafner and Nadler 1988), the pocket gophers (family Geomyidae) have much longer generation times and a slower substitution rate, in comparison with their chewing lice parasites (Phthiraptera: Ischnocera) (Light 2007). The gophers do not evolve detectably in response to phenotypic changes in the lice (which live on and in the gophers' fur), but the lice have been shown to have an evolutionary history that is highly congruent with that of the gophers. This congruence has been shown to be statistically significant in many studies (Hafner and Nadler 1988, Hafner and Page 1995, Demastes and Hafner 1993 and others). The smaller, evolutionarily more agile lice (in the sense of this faster mutation rate and other potential contributing factors) are coevolving with their changing environment, i.e., their host species, as it diverges and diversifies.

Assume that we are given the phylogenetic tree, H, for the "host" and P, for the "parasite" (or "pathogen") species as well as the association, or "mapping," between extant parasites and their hosts (i.e., an association between the tips, or leaves, of the two trees). Also assume for now that we have some way of measuring congruence between trees. The most fundamental question that we can ask is "how does the level of congruence between these two trees compare to random data?" If the level of congruence is statistically higher than that of random trees (or the same trees but with randomized associations of the tips), then we can reject the null hypothesis that the similarity of these trees is due to chance. In addition, we can ask questions such as:

1. What were the associations of the (hypothetical) ancestral species in P to those in H?
2. How long have the species in H and P been interacting with each other?

3. What is the level of cospeciation between *H* and *P*?
4. How often do parasites jump species boundaries to new hosts in *H*?
5. Are there particular branches in either tree that appear to contribute to most to the (in)congruence between *H* and *P*?

Collectively, these and related questions can be called the "cophylogeny problem." These questions require computational methods, and these methods have become increasingly sophisticated over the last two decades. This chapter examines these methods, their underlying assumptions, and their strengths and weaknesses.

20.2 An Overview of Cophylogeny Methods

The "cophylogeny problem" is relatively new in biology, partly because it is computationally difficult to solve, but also because there have been, up until the present, very few good datasets to which to apply these kinds of analyses. While it is common to find evolutionary trees for groups of species in the scientific literature, there have been relatively few studies in which trees for two or more ecologically linked groups of species have been estimated.

We also require some knowledge about the associations between the tips of both trees, which often requires substantial fieldwork. Often, the predominant form of observed association between hosts and parasites (or pathogens) is in the form of "presence" that a particular parasite is found on a particular host, or "absence," it was not found on that host. *Host specificity*—the degree to which a parasite prefers one host to another—is therefore usually judged simply by "absence" of observation on other hosts, which is a clear statistical bias toward favoring false negatives. Indeed, there is consistent evidence that across many systems, parasites or pathogens are limited more by opportunity than by common descent (e.g., Bush and Clayton 2006; Poulin 2007; Poulin and Keeney 2008).

So our data have been limited quite significantly because they are hard to get, and often, in particular in light of this host specificity issue, they are also hard to interpret. We also note the unavoidable effect of variable intensity of study of different host/parasite systems, which is that more intensely studied groups are much more likely to have knowledge about their associations gathered than are less well-studied groups. This must lead to missing associations, in turn leading to less well-supported hypotheses of codivergence or coevolution.

The first computational method for measuring the congruence of host and parasite trees given their tip associations was Brooks' Parsimony Analysis (BPA) (Brooks 1981; Brooks et al. 2001). BPA first codes the parasite tree *P* as a set of binary vectors with one vector per tip. Next, each tip of host tree *H* is assigned the vector of its associated tip in *P*. Those vectors are interpreted as character vectors, and the well-known Fitch Parsimony Algorithm (Fitch 1971; Felsenstein 2004) is used to find the minimum number of state changes required to transform an all-zero

binary vector at the root into the binary vectors at the tips. The minimum number of state changes is interpreted as a measure of congruence between P and H.
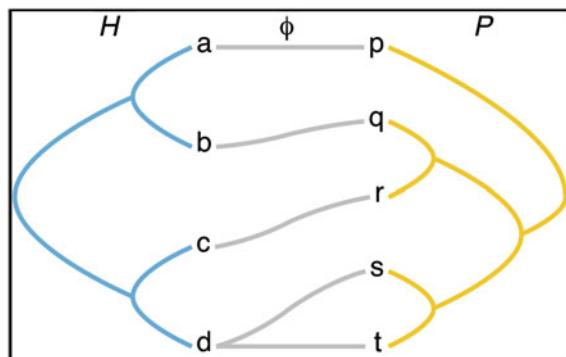
The BPA method suffers from a lack of a natural interpretation of how the computed score corresponds to evolutionary events. The merits of BPA have been questioned, while Brooks has vehemently defended his approach, resulting in an occasionally healthy debate (Page 1990; Ronquist and Nylin 1990; Brooks et al. 2001). No known publicly distributed software packages implement BPA.

Another approach, called ParaFit (Legendre et al. 2002), codes each of two trees P and H and the tip associations as a separate matrix. These three matrices are multiplied in a specific order, and one of several statistics can be computed on the result to score the congruence of the host and parasite trees. This method is useful for statistical analyses of congruence, but does not offer insights on where the two trees differ and the evolutionary events that might explain those differences—although it can be used to identify which tip associations are contributing most to the measured congruence (Legendre et al. 2002). The AxParaFit software tool (Stamatakis et al. 2007) is an optimized version of ParaFit, and CopyCat is a software tool that provides a convenient user interface for AxParaFit (Meier-Kolthoff et al. 2007).

The congruence index method (de Vienne et al. 2007) computes another measure of similarity between each of two trees P and H by determining the largest subtree common to both trees. Specifically, a subtree of P (or H) is obtained by removing tips and then collapsing internal nodes with only one child. A subtree T is said to be a *maximal agreement subtree* if it can be obtained by performing this process from P and from H and there exists no larger subtree that can be constructed in this way. The number of tips in a maximum agreement subtree is used as the measure of congruence between the two trees. This measure can be compared to those of randomly generated trees of the same size as P and H in order to test the null hypothesis that the two trees are congruent by chance. While some concerns have been raised about this method (Kupczok and von Haeseler 2009), the authors note that this test is intended only to "get a first and rapid (computationally cheap) insight on the topological congruence between two trees" before proceeding to more detailed analyses (de Vienne et al. 2009).

All of the aforementioned methods provide statistical measures of congruence between pairs of trees, but do not attempt to infer the best supported set of events that explain their congruence and incongruence. In contrast, *event-based reconstruction methods* provide *both* congruence scores *and* corresponding mappings, or *reconstructions*, of P into H. For this reason, event-based methods are said to solve the *cophylogeny reconstruction problem*. Event-based based methods offer rich insights into the evolutionary histories of pairs of phylogenies, and for that reason, we focus on these methods in the remainder of this chapter. Among these methods and software tools, which are discussed in more detail in the next section, are TreeFitter (Ronquist 1995), TreeMap (Charleston and Page 2002), CoRe-PA (Merkle et al. 2010), and Jane (Conow et al. 2010).

Fig. 20.1 A simple tanglegram

The input to the cophylogeny reconstruction problem is a triple (H, P, ϕ) comprising a host tree H, a parasite tree P, and a function ϕ that associates each parasite tip with a host tip. The triple (H, P, ϕ) can be represented graphically as a *tanglegram* as shown in the example in Fig. 20.1. The objective is to map P into H to associate ancestral nodes of P with locations in H. Host and parasite tips are associated as specified by ϕ and each ancestral association places a node p in P on a node or an edge of H, implying a set of coevolutionary processes, or *events*, that reconcile the two trees.

The possible events are *cospeciation*, *duplication*, *lineage sorting*, and *host switching*. Cospeciation corresponds to contemporaneous speciation in the host and parasite (mapping a parasite tree node onto a host tree node), duplication corresponds to speciation in the parasite that is not contemporaneous with speciation in the host (mapping a parasite tree node onto a host tree edge), lineage sorting corresponds to a parasite lineage failing to speciate with a host (an edge of the parasite tree that passes through a node in the host tree), and host switching corresponds to a duplication event in which one of the two parasite lineages "jumps" to another part of the host tree.

Assume that a cost is associated with each of the four types of events: cospeciation, duplication, lineage sorting, and host switching. Our objective is to find a reconstruction of P into H that is consistent with the tip mapping ϕ and that minimizes the total cost of the events in the mapping. For example, Fig. 20.2 shows three different reconstructions for the trees P and H in Fig. 20.1. If cospeciation has cost 0 and all other event costs have cost 1, then the cost of the reconstruction in Fig. 20.2a–c is 3, 7, and 4, respectively. The choice of event costs is a notoriously difficult problem and is addressed in more detail in the next section.

We note that the cophylogeny reconstruction problem arises in several contexts other than host–parasite evolution. For example, a fundamental problem in phylogenomics is that of reconciling gene trees and species trees. In the widely studied DTL (Duplication–Loss–Transfer) model, the four events are speciation, duplication, loss, and horizontal gene transfer which, mathematically, correspond to the cospeciation, duplication, lineage sorting, and host switching, respectively.
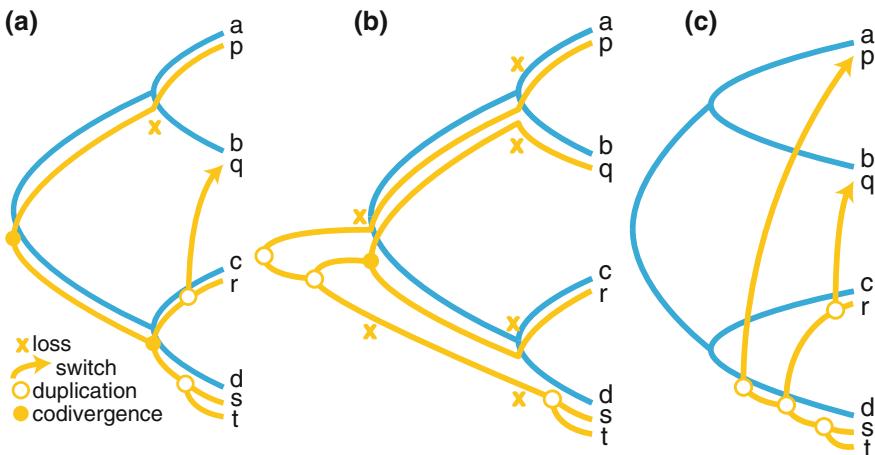
Fig. 20.2 Three possible reconstructions of the parasite tree P (*orange*) into the host tree H (*blue*) for the tanglegram in Fig. 20.1. The first **a** is an optimal map if, for example, duplication, host switch, and loss events each cost 1 and cospeciation costs 0 and assuming that we have no further information about the relative timing of events between *P* and *H*. **b** The second is optimal if host switches are prohibitively expensive and suggests a history dominated by ancient duplications and recent losses. The last solution, **c** would be optimal if we knew that all the divergence events on *P* were very recent, such as the case when considering viruses switching between host species

Similarly, biogeographical analyses of species and their habitats require reconstruction of species trees and their area cladograms interacting through vicariance, sympatric speciation, dispersal, and extinction which are again analogous to the four cophylogenetic events. While these domains are different, the events are mathematically analogous, and thus, the reconstruction problems are the same. Thus, the cophylogeny reconstruction problem is broadly applicable to a number of domains beyond host–parasite coevolution.

20.3 Computational Complexity, Algorithms, and Heuristics

The problem of finding minimum cost reconstructions under the four event types (cospeciation, duplication, lineage sorting, and host switching) is known to be computationally intractable or, in the parlance of computational complexity theory, NP-complete (Ovadia et al. 2011). Roughly, this means that while the problem can be solved by algorithms whose running time grows exponentially with the number of tips in the trees, there is no known efficient (polynomial-time) algorithm for the problem. Moreover, complexity theory provides strong evidence that no efficient algorithm exists for any NP-complete problem.

Thus, two approaches can be taken: *Heuristics* can be used to quickly find good, but not necessarily optimal, solutions, or *exact algorithms* can be used to find solutions of optimal cost, but these algorithms have exponential running times that are prohibitive for large trees.

Exact solutions to the cophylogeny reconstruction problem can, however, be found in polynomial time if host switching is prohibited. For this reason, some early event-based approaches simply prohibited host switching. However, host switching is known to occur in many biological systems and thus excluding this event severely limits the scope of the analyses that can be conducted.

Exact solutions can also be found in polynomial time, even with host switches, if the tree is fully dated—that is, if a total ordering is given on the relative times of the internal nodes (Libeskind-Hadas and Charleston 2009; Doyon et al. 2010). However, accurately dating the internal nodes of a phylogenetic tree is generally difficult (Rutschmann 2006). In the absence of reliable dates, the problem becomes computationally intractable because of the difficulty of maintaining time consistency. If, however, the time consistency constraint is relaxed, potentially permitting reconstructions that require contradictory orderings on the relative times of events, then the problem can also be solved by efficient polynomial-time algorithms. Empirical results suggest that these algorithms rarely introduce timing inconsistencies (Addario-Berry et al. 2003). Moreover, timing inconsistencies can be easily detected. However, if a solution is found to have timing inconsistencies, this approach offers no recourse.

While maximum parsimony reconstructions attempt to infer evolutionary events, the reconstructions and their numerical scores alone do not suffice to address the question of whether or not two trees are congruent. Congruence is usually determined by performing random permutation tests. Specifically, the parasite tree P or the association ϕ between host and parasite tips is randomized some number of times and the score is recomputed for each randomized instance. If x % or less of random trials do not get as good a score as the original P, then we can say P is significantly congruent at the x % level. That is, x (or more accurately x/100) serves as an empirical p-value. Since these tests typically comprise hundreds or even thousands of trials, the efficiency of the reconstruction algorithms is particularly important.

20.4 Software

The prevailing software tools for the cophylogeny reconstruction problem are TreeFitter (Ronquist 1995), TreeMap (Charleston and Page 2002;Charleston 2012), CoRe-PA (Merkle et al. 2010), and Jane 4 (Conow et al. 2010). TreeFitter and the original version of TreeMap (TreeMap 1) were among the earliest tools. While TreeFitter is no longer supported by its developer, it is available as an open source project. CoRe-PA (based on a predecessor called Tarzan) and Jane are more

recent additions to this small jungle[1] of tools. This section describes the underlying approaches and major features of these different tools. While version numbers are henceforth omitted, our comparisons refer to TreeMap 3, CoRe-PA 0.5.1, and Jane 4. The URLs for these tools are available from the Online Practical Material for this chapter (http://www.mpcm-evolution.com).

TreeMap uses an exact algorithm to find optimalreconstructions. In addition, it can find multiple so-called *Pareto-optimal* solutions that are optimal for different event costs. The issue of event costs and Pareto optimality is explored in more detail in the next section. Since the running time of exact algorithms grows exponentially with the number of tips in the trees, TreeMap offers mechanisms for reducing running time by limiting the number of host switch events that it considers.

TreeFitter uses two different algorithms. Its "lower-bound" algorithm efficiently finds reconciliations of optimal cost by relaxing the time consistency constraint, possibly resulting in invalid solutions. Its "upper-bound" algorithm is not documented but evidently finds optimal time-consistent reconstructions, and this implies that its running time is exponential. CoRe-PA also relaxes the time consistency constraint and thus efficiently finds solutions that may be invalid.

Jane exploits the fact that optimal cost reconciliations can be found efficiently if the trees are timed. If the trees are not timed, Jane explores a sample of the possible relative times of the events and solves each of them efficiently, reporting the best solutions it finds. Thus, Jane's solutions are always time consistent, guaranteed to be optimal for timed trees, but are not guaranteed to be of optimal cost for untimed trees.

Thus, in terms of running time for untimed trees, CoRe-PA is the fastest, Jane is intermediate, and TreeMap is the slowest. (TreeFitter is fast in the "lower-bound" mode and slow in the "upper-bound" mode.) Conversely, TreeMap finds solutions that are guaranteed to be valid and of optimal cost (assuming the number of host switches is not constrained), Jane finds solutions that are guaranteed to be valid and have good but not necessarily optimal cost, and CoRe-PA and TreeFitter ("lower-bound mode") find solutions that may not be valid, but if they are, then their cost is optimal.

Beyond the algorithmic similarities and differences mentioned above, these tools share a number of common features. All of these tools allow for user-defined event costs, compute reconciliations, and have functionality for performing randomized permutation tests. TreeFitter, CoRe-PA, and Jane provide methods for handling polytomous trees (discussed further in the next section). Additionally, TreeMap, CoRe-PA, and Jane all offer graphical user interfaces and tools for editing pairs of trees and their tip assignments and viewing the reconstructions.

These tools also have unique features and limitations. TreeFitter has a command-line interface but no graphical user interface and does not render the

---

[1] Indeed, the jungle metaphor originates from a data structure called a "jungle" which is used in the TreeMap tool.

reconstructions. Recognizing that the choice of event costs can affect the reconstructions, TreeFitter can explore a user-specified range of event costs in user-specified increments. However, it also imposes constraints on the permitted relationships of event costs.

A unique feature of CoRe-PA is that it offers a mode that attempts to automatically infer event costs based on the assumption that these costs should be inversely proportional to the frequency of their corresponding events. CoRe-PA's "parameter-adaptive" approach, therefore, uses a mathematical optimization technique to find event costs that are inversely proportional to the frequency of events in the maximum parsimony reconstructions that they induce. This allows the user to avoid the difficult and ill-defined process of estimating appropriate event costs, but it is not clear that the inferred costs are necessarily biologically realistic or give rise to the most plausible reconstructions.

Jane supports preferential host switching (Charleston and Page 2002), which permits the user to assign different costs for host switch events, depending on the distance from the original host to the new host. It also permits multi-host (i.e., widespread) parasites using an event called "failure to diverge" (discussed in more detail in the next section). Additionally, it has a "time zone" feature that allows complete or partial annotations on relative times of events and finds solutions that respect the provided timing information.

20.5 Computational and Biological Issues

This section examines some of the computational and biological issues in cophylogenetic reconstruction in more detail.

20.5.1 Performance Optimizations

As datasets get larger, there is a need for faster heuristics. One promising approach is to "condense" a problem instance by recognizing recurring patterns in the two trees and collapsing them to create a slightly smaller problem instance. Rather than optimizing a score, this approach (Drinkwater and Charleston, in prep) is built purely for speed and makes no guarantees on optimality. It does produce a reconstruction, but it is as a result of the collapsing process. This algorithm is even faster than the best dynamic programming algorithms for reconstruction discovered to date. Its accuracy is commendable despite having no guarantees: In the majority (68) of 102 published tanglegrams, it was shown to find solutions as good as those found by Jane.

20.5.2 Event Costs

One way that event-based approaches differ from one another is in the way that they address event costs. Generally, these costs are assumed to satisfy the following weak constraint: Cospeciation cost (c_c) is strictly less than duplication (c_d), host switch (c_w), or loss (c_x). These have been chosen after the work of Fahrenholz (1913) who famously stated that parasite phylogeny should mimic host phylogeny.

It is very difficult to know what these event costs should really be though: Every biological system is different, and what works in one case will not be appropriate for another. It should be noted that the event costs are "unit-less" in the sense that only their relative values are important: Scaling all event costs by any constant factor simply scales the total cost of the solution by that value. Generally speaking, cospeciation is assumed to have a low cost (e.g., 0 or some small positive value) and all other events have a larger positive cost. In some analyses, cospeciation is set to have a cost of -1 and all other events have a cost of 0 so that minimum cost reconstructions maximize the number of cospeciations.

Some approaches such as TreeFitter and Jane require the user to specify the event costs or range of possible event costs, while others, such as CoRe-PA, attempt to infer event costs automatically. In contrast, TreeMap does not use explicit event costs but rather seeks to find all solutions that are optimal for some choice of event costs. Specifically, we can represent a reconstruction by an *event count vector* (c, d, t, s) which counts the number of cospeciations, duplications, transfers, and host switches, respectively. An event count vector $v = (c, d, t, s)$ is said to be *strictly better* than another event count vector $v' = (c', d', t', s')$ if each entry in vector v is less than or equal to its corresponding entry in vector v' and at least one entry in v is less than its corresponding entry in v'. A reconstruction with an event count vector v is said to be Pareto optimal if there exists no other reconstruction with an event count vector strictly better than v. We use the term "Pareto optimal" to refer to both the reconstruction and its event count vector.

TreeMap was the first reconstruction algorithm to use the notion of Pareto-optimal reconstructions. Since the underlying algorithm in TreeMap has exponential worst-case time, in practice it finds a subset of the Pareto-optimal event count vectors. Moreover, even for a single set of event costs, there may be a large number of different equally good maximum parsimony reconciliations. And, in the case of searching for a Pareto front, there will be an even larger number of candidate reconstructions. Thus, it is not computationally feasible to enumerate all of the distinct maximum parsimony reconciliations.

Recently, Libeskind-Hadas et al. have devised efficient polynomial-time algorithms that compute all Pareto-optimal event count vectors and determine the number of distinct reconstructions for each Pareto-optimal event count vector (Libeskind-Hadas 2013). These algorithms are implemented in a suite of tools called *x*scape. For example, the costscape tool lists the set of all Pareto-optimal event count vectors and partitions the cost space into "regions" such that all event costs in a given region give rise to the same set of Pareto-optimal reconciliations.

Figure 20.3 shows the regions found by costscape for a dataset of figs and the fig wasps that pollinate them (Weiblen and Bush 2002) where cospeciation is fixed to cost 0, duplication is normalized to cost 1, and loss and transfer (switch) have costs ranging from 0.2 to 5, relative to the unit cost of duplication. Each color-coded region corresponds to the subset of the cost space with identical maximum parsimony solutions and those solutions have Pareto-optimal event count vectors as shown in the legend. In addition, the legend shows the number of distinct maximum parsimony reconciliations in that region. The point corresponding to Jane's default event costs is also shown, demonstrating that a single event cost provides only a small snapshot of the totality of plausible solutions.

20.5.3   Event Support
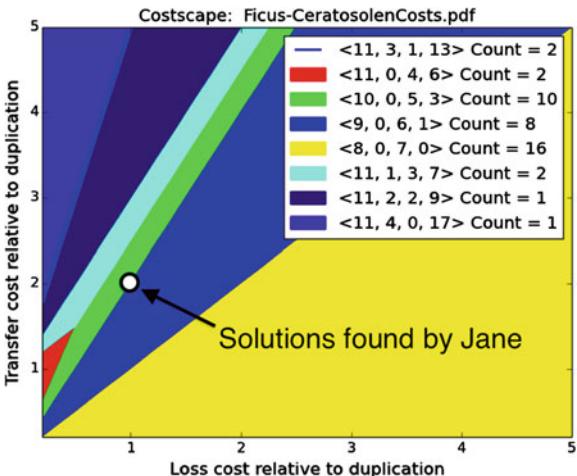
We have noted earlier that maximum parsimony reconciliation can produce a large number of equally optimal solutions for a single set of event costs and an even larger number over a range of event costs. It is important, therefore, to distinguish between events that are robust (e.g., occur in many solutions and are insensitive to perturbations of event cost) and those that are not. Pareto-optimal reconciliation offers new and promising approaches to identifying highly supported events.

In general, the number of different event cost regions can grow quadratically and the number of distinct reconciliations can grow exponentially with the size of the parasite tree. Nonetheless, we can count the number of reconciliations and enumerate the events in those solutions in polynomial time. The number of reconciliations grows exponentially because of the number of ways of choosing the events that comprise the reconciliation. So, the number of events is polynomially bounded even though the number of reconciliations is not. This observation turns out to be useful in identifying strongly supported events.

Given a range of costs on the transfer and loss events, we find the Pareto-optimal regions, and in the process, in each region, we collect the set of events that are shared by every reconstruction in that region. Those events are inferred to be strongly supported for that region. We can define an event to have α-consensus support, $0 < \alpha \leq 1$, if the event arises in every reconstruction in a fraction of at least α of the regions. Two special cases are majority consensus (events that arise in more than half of all regions) and strict consensus (events that arise in all regions). This is in contrast to recent work that defines event support based on a single fixed cost for the events (Bansal et al. 2013; Nguyen et al. 2013).

These event consensus support values can be computed by efficient polynomial-time algorithms implemented in the eventscape tool (Libeskind-Hadas 2013). For example, Fig. 20.4 shows a summary of results obtained applying eventscape to five host–parasite datasets: figs and fig wasps (Weiblen and Bush 2002), gophers and lice (Hafner and Nadler 1998), indigobirds and finches (Sorenson et al. 2004), seabirds and lice (Paterson et al. 2000), and trees and moths (Kawakita et al. 2004). One cost space had transfer and loss costs ranging from 0.5 to 2, a 4-fold

Fig. 20.3 The Pareto-optimal event count vectors for the fig wasp dataset and the corresponding regions. The cost space ranges from 0.2 to 5 for both transfers and losses. The solutions found using Jane with its default cost values are indicated with a *black hollow dot*. Note also that the event count vector $\langle 11, 3, 1, 13 \rangle$ has a region that is a line rather than a *polygon*

range, and the second had transfer and loss costs ranging from 0.2 to 5, a 25-fold range. For each dataset and range value (4 or 25), we counted the number of events of each type supported at the given level and normalized by the total number of events of that type found in all regions in the cost space.

These results show that even under strict consensus, a substantial fraction of events can be identified that arise in every reconciliation across the cost space. Interestingly, cospeciation and transfer events appear to be more highly supported than duplication and loss events.

20.5.4 Multi-host Parasites

Generally, the cophylogeny problem assumes that each tip in the parasite tree is only associated with a single host tip (i.e., ϕ is a function from parasite tips to host tips). This is a significant limitation because there are many cases in which parasites associate with multiple hosts. To date, there are no generally accepted methods of dealing with the multiple-host problem.

Suppose one parasite p is associated with multiple host tips. If these hosts are all each others' closest relatives in the host tree, then this set of hosts constitutes a *monophyletic group*, and we can simply denote this clade as a single taxonomic unit and consider it as a single "meta-tip" in our analysis (Fig. 20.3a). However, if the hosts of p are spread across the host tree more widely (as in Fig. 20.3b), then this approach will not work. We need some way of accounting for the (possibly myriad) alternatives for the history of associations of p with sets of hosts within H, at least back until the common ancestor of these widespread hosts, and possibly beyond (Figs. 20.3c, 20.5).
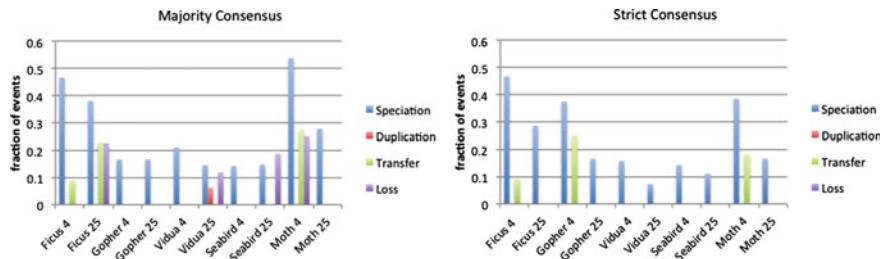
Fig. 20.4 Fraction of events of each type found under majority and strict consensus for five host–parasite datasets, each using a 4-fold and 25-fold range of event costs
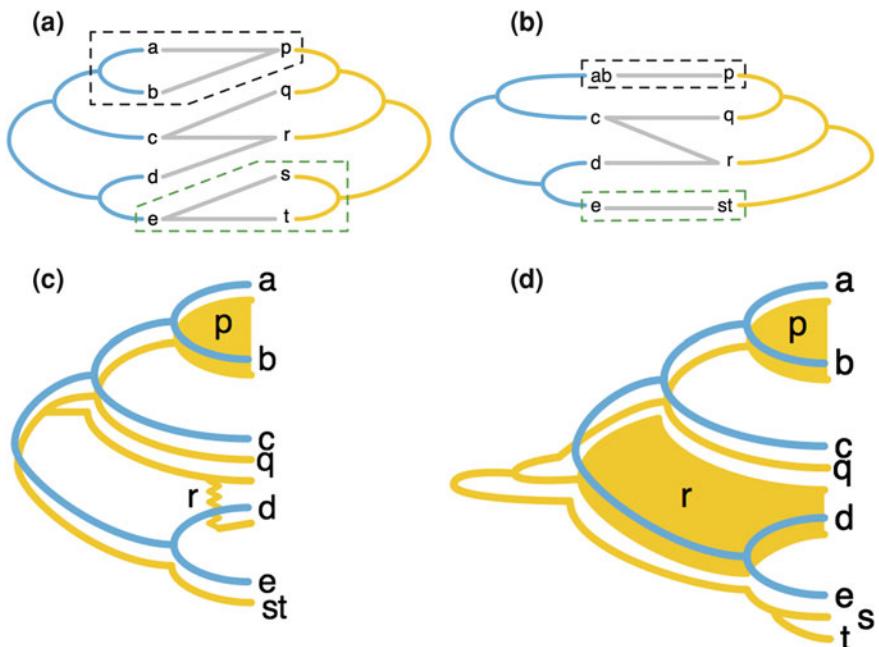
Fig. 20.5 a A tanglegram showing parasites with multiple hosts and hosts with multiple parasites. Parasite p has multiple hosts but it is on a single monophyletic clade which is then collapsed to a single OTU in the same subfigure, **b** a case in which parasite r's hosts do not form such a clade and cannot be collapsed in this way, and **c, d** two histories showing multi-locations of ancestors of p

New types of "events" are required to explain widespread host associations. One such event, called *failure to (co)diverge*, allows a parasite lineage to "split" when its associated ancestral host lineage speciates as shown in Fig. 20.4a. Association with widespread hosts can be explained, for example, by a series of failure to diverge events beginning at the most recent common ancestor of those

hosts as shown in Fig. 20.4b. Some systems, such as Jane, support failure to diverge events and use the "most recent common ancestor solution."

Another type of event that can be used to explain multi-host parasites might be called *spreading*. In contrast to failure to diverge, spreading allows a parasite lineage to on its current host lineage and sends a copy to another host lineage on a different part of the tree. This is shown in Fig. 20.4a. Currently, no cophylogeny mapping system supports spreading events, though some promising approaches are currently being studied by the authors and others.

20.5.5 Polytomous Trees

Phylogenetic trees often contain non-binary branches, known as *polytomies* or *multifurcations*. These polytomies may be artifacts of the phylogenetic inference process (*soft polytomies*) or biologically meaningful (*hard polytomies*). Polytomies create additional challenges for cophylogenetic reconstruction.

When both the host and parasite trees contain a polytomy in corresponding locations, it may seem natural to resolve those polytomies identically. However, this is a highly dubious approach as it risks serious bias toward finding significant congruence based on that resolution. Moreover, in many cases, the polytomies in the two trees do not occur in corresponding locations. For example, one tree might have a polytomy at the root, while the other tree is fully resolved at the root.

TreeFitter, CoRe-PA, and Jane all handle polytomous trees in different ways. TreeFitter resolves polytomous trees by constructing a user-defined number of randomly resolved trees and performs reconciliations using each of these trees. CoRe-PA does not specify how polytomies are handled, and Jane uses a population of randomly resolved trees, allowing the user to specify whether the polytomy resolutions should be interpreted as occurring in rapid succession or may be temporally interleaved with other events in the tree.

Finally, we provide a worked-out example of a cophylogenetic analysis using both TreeMap and Jane in the Online Practical Material (http://www.mpcm-evolution.com).

20.6 Conclusions

The techniques and tools described in this chapter have been applied to a wide set of biological systems. But challenges remain in dealing with the messiness of data in which the trees are not fully known and possibly not even treelike (e.g., reticulate phylogenies due to hybridization), associations are blurry, and there may even be multiple interacting sets of organisms such as in the highly complex fig/fig wasp/nematode/wolbachia system (Shoemaker et al. 2002, Jackson 2004). Additional open problems remain in dealing with multi-host parasites, polytomies,

determining "support values" for events in reconciliations, among others. Fortunately, there is an active research community pursuing these problems and we expect that increasingly sophisticated techniques and tools will be available in years to come.

References

Addario-Berry L, Hallett M, Lagergren J (2003) Towards Identifying Lateral Gene Transfer Events. Pacific Symposium on Biocomputing 8:279–290

Bansal MS, Alm EJ, Kellis, M. (2013) Reconciliation revisited: handling multiple optima when reconciling with duplication, transfer, and loss. In: Research in Computational Molecular Biology (pp. 1–13). Springer, Berlin Heidelberg

Brooks DR (1981) Hennig's parasitological method: a proposed solution. Syst Zool 30:229–249

Brooks DR, van Veller MGP, McLennan DA (2001) How to do BPA, really. J Biogeogr 28(3):345–358

Bush SE, Clayton DH (2006) The role of body size in host specificity: reciprocal transfer experiments with feather lice. Evolution 60(10):2158–2167

Charleston MA (1998) Jungles: a new solution to the host-parasite phylogeny problem. Math Biosci 149:191–223

Charleston MA (2012) TreeMap 3b. A Java program for cophylogeny mapping.http://www.sydney.edu.au/engineering/it/∼mcharles/software/treemap/

Charleston MA, Page RDM (2002) TreeMap 2. A Macintosh program for cophylogeny mapping. http://www.sydney.edu.au/engineering/it/∼mcharles/software/treemap/

Charleston MA, Robertson DL (2002) Preferential host switching by primate lentiviruses can account for phylogenetic similarity with the primate phylogeny. Syst Biol 51(3):528–535

Chun TW, Davey RT, Engel D, Lane HC, Fauci AS (1999) AIDS: re-emergence of HIV after stopping therapy. Nature 401:874–875

Conow C, Fielder D, Ovadia Y, Libeskind-Hadas R (2010) Jane: a new tool for the cophylogeny reconstruction problem. Algorithms Mol Biol 5:16. doi:10.1186/1748-7188-5-16

Cuthill JH, Charleston M (2012) Phylogenetic codivergence supports coevolution of mimetic Heliconius butterflies. PLoS ONE 7(5):e36464

Dobzhansky T (1973) Nothing in biology makes sense except in the light of evolution. Am Biol Teacher vol 35

de Vienne DM, Giraud T, Martin OC (2007) A congruence index for testing topological similarity between trees. Bioinformatics 23(23):3119–3124

de Vienne DM, Giraud T, Martin OC (2009) In response to comment on 'a congruence index for testing topological similarity between trees'. Bioinformatics 25(1):150–151

Demastes JW, Hafner MS (1993) Cospeciation of pocket gophers (Geomys) and their chewing lice (Geomydoecus). J Mammal 74(3):521–530

Doyon J-P, Scornavacca C, Gorbunov KY, Szöllosi GJ, Ranwez V and Berry V (2010) An Efficient Algorithm for Gene/Species Trees Parsimonious Reconciliation with Losses, Duplications and Transfers. RECOMB-CG 93–108

Fahrenholz H (1913) Ectoparasiten und Abstammungslehre. Zoologischer Anz 41:371–374

Felenstein J (2004) Inferring phylogenies (Vol 2) Sinauer Associates, Sunderland MA

Fitch WM (1971) Toward defining the course of evolution: minimum change for a specified tree topology. Syst Zool 20(4):406–416

Hafner MS, Nadler SA (1988) Phylogenetic trees support the coevolution of parasites and their hosts. Nature 332:258–259

Hafner MS, Page RDM (1995) Molecular phylogenies and host-parasite cospeciation: gophers and lice as a model system. Philos Trans Roy Soc London (B) 349(1327):77–83

Jackson AP (2004) Cophylogeny of the ficus microcosm. Biol Rev Camb Philos Soc 79:751–768

Kawakita A, Takimura A, Terachi T, Sota T, Kato M (2004) Cospeciation analysis of an obligate pollination mutualism: Haveglochidion trees (euphorbiaceae) and pollinating epicephala moths (gracillariidae) diversified in parallel? Evolution 58(10):2201–2214

Kupczok A, von Haeseler A (2009) Comment on 'a congruence index for testing topological similarity between trees'. Bioinformatics 25(1):147–149

Legendre P, Desdevises Y, Bazin Y (2002) A statistical test for host-parasite coevolution. Syst Biol 51(2):217–234

Libeskind-Hadas R, Charleston MA (2009) On the Computational Complexity of the Reticulate Cophylogeny Reconstruction Problem. Jour Comput Biol 16(1):105–117

Libeskind-Hadas, R, Wu Y-C, Bansal MS, Kellis M (2013) Pareto-optimal Phylogenetic Tree Reconciliation, in Proceedings of ISMB 2014

Light JE, Hafner MS (2007) Cophylogeny and disparate rates of evolution in sympatric lineages of chewing lice on pocket gophers. Mol Phylogenet Evol 45(3):997–1013

Meier-Kolthoff JP, Auch AF, Huson DH, Göker M (2007) **COPYCAT**: co-phylogenetic analysis tool. Bioinformatics 23(7):898–900

Merkle D, Middendorf M, Wieseke N (2010) A parameter-adaptive dynamic programming approach for inferring cophylogenies. BMC Bioinf 11(Suppl 1):S60

Nguyen TH, Ranwez V, Berry V, Scornavacca C (2013) Support measures to estimate the reliability of evolutionary events predicted by reconciliation methods. PLoS ONE 8(10):e73667

Ovadia Y, Fielder D, Conow C, Libeskind-Hadas R (2011) The cophylogeny reconstruction problem is NP-complete. J Comput Biol Jan 18(1):59–65

Page RDM (1990) Component analysis: a valiant failure? Cladistics 6:119–136

Paterson AM, Wallis GP, Wallis LJ, Gray RD (2000) Seabird and louse coevolution: complex histories revealed by 12S rRNA sequences and reconciliation analyses. Syst Biol 49(3):383–399

Poulin R (2007). Evolutionary ecology of parasites. Edn 2. Princeton University Press, Princeton NJ

Poulin R, Keeney DB (2008) Host specificity under molecular and experimental scrutiny. Trends Parasitol 24:24–28

Ronquist F, Nylin S (1990) Process and pattern in the evolution of species associations. Syst Zool 39(4):323–344

Ronquist F (1995) Reconstructing the history of host-parasite associations using generalized parsimony. Cladistics 10(1):73–89

Rutschmann F (2006) Molecular dating of phylogenetic trees: a brief review of current methods that estimate divergence times. Diversity and Distributions 12(1):35–48

Shoemaker DD, Machado CA, Molbo D, Werren JH, Windsor DM, Herre EA (2002) The distribution of Wolbachia in fig wasps: correlations with host phylogeny, ecology and population structure. Proc R Soc London (B) 269:2257–2267

Siddall ME (2005) Bracing for another decade of deception: the promise of secondary Brooks' parsimony analysis. Cladistics 21:90–99

Sorenson MD, Balakrishnan CN, Payne RB (2004) Clade-limited colonization in brood parasitic finches (Vidua spp.). Syst Biol 53(1):140–153

Stamatakis A, Auch AF, Meier-Kolthoff JP, Göker M (2007) AxPcoords & parallel AxParafit: statistical co-phylogenetic analyses on thousands of taxa. BMC Bioinf 8:405

Thompson JN (1994) The coevolutionary process. University of Chicago Press, Chicago

Weiblen GD, Bush GW (2002) Speciation in fig pollinators and parasites. Mol Ecol 11:1573–1578

Chapter 21
Phylogenetic Prediction to Identify "Evolutionary Singularities"

Charles L. Nunn and Li Zhu

Abstract Understanding adaptive patterns is especially difficult in the case of "evolutionary singularities," i.e., traits that evolved in only one lineage in the clade of interest. New methods are needed to integrate our understanding of general phenotypic correlations and convergence within a clade when examining a single lineage in that clade. Here, we develop and apply a new method to investigate change along a single branch of an evolutionary tree; this method can be applied to any branch on a phylogeny, typically focusing on an a priori hypothesis for "exceptional evolution" along particular branches, for example in humans relative to other primates. Specifically, we use phylogenetic methods to predict trait values for a tip on the phylogeny based on a statistical (regression) model, phylogenetic signal (λ), and evolutionary relationships among species in the clade. We can then evaluate whether the observed value departs from the predicted value. We provide two worked examples in human evolution using original R scripts that implement this concept in a Bayesian framework. We also provide simulations that investigate the statistical validity of the approach. While multiple approaches can and should be used to investigate singularities in an evolutionary context—including studies of the rate of phenotypic change along a branch—our Bayesian approach provides a way to place confidence on the predicted values in light of uncertainty about the underlying evolutionary and statistical parameters.

Convergence is fundamental to the comparative approach to testing adaptive hypotheses in biology. In short, we can be more confident that a trait is an adaptation if it has evolved repeatedly—rather than once—in association with another trait, environment, or other factors (Pagel 1994). Phylogeny is essential to

C. L. Nunn (✉)
Department of Evolutionary Anthropology, Duke University,
Durham, NC 27514, USA
e-mail: charleslnunn@gmail.com

L. Zhu
Department of Statistics, Harvard University, Cambridge, MA 02138, USA

this endeavor because an evolutionary tree provides the scaffolding upon which to identify evolutionary origins of traits and their covarying factors. Thus, phylogenetic methods are widely used to identify correlated trait evolution, to probe the factors that drive speciation and extinction, and to estimate rates of evolutionary change (Harvey and Pagel 1991; Garland et al. 2005; Nee 2006; Maddison et al. 2007; Martins 1994; Nunn 2011).

The convergence approach has proved incredibly powerful, yet this approach is not appropriate for investigating evolutionary singularities—i.e., traits that evolved in only one lineage in the clade of interest. Such a trait may appear in a single taxon on the tree, and thus, the evolutionary event occurred on the branch leading to that taxon (autapomorphy), or the singularity may occur on an internal branch, leading to its representation in all species in a subclade of multiple species (a synapomorphy). As with many concepts in cladistics, identifying singularities depends on the taxonomic level under consideration. Thus, we might say that "winged flight" is an evolutionary singularity in mammals (i.e., bats), but not among vertebrates more broadly (i.e., bats and birds).

Evolutionary singularities are especially relevant in studies of human evolution. Indeed, humans are unusual mammals with a suite of "zoologically unprecedented capacities" (Tooby and DeVore 1987, p. 183), such as language, walking with a striding gait, and wearing clothing. Humans possess complex cultural traits that build on other cultural traits and thus exhibit cumulative cultural evolution (Tennie et al. 2009). In terms of quantitative traits, humans have relatively large brains and exhibit longer periods of parental care than are found in other primates of our body mass. These traits can be examined in broad phylogenetic context using the comparative method (Barton 1996; Dunbar 1993; Deaner et al. 2000). Evolutionary anthropologists are interested in identifying the characteristics of humans that make us unique relative to other primates (Martin 2002; Kappeler and Silk 2009; Rodseth et al. 1991). Yet the novelty of our traits makes it challenging—some might say impossible—to quantitatively investigate the factors that influenced their evolution using convergence-based comparative approaches.

Considering quantitative characters such as brain size or body mass, two related approaches can be used to investigate evolutionary singularities. One approach "predicts" trait values for tips of the tree and quantifies deviations from this prediction (Garland and Ives 2000; Nunn 2011; Organ et al. 2011). The other approach estimates rates of evolutionary change along the branch of interest and then compares this rate to other branches on the tree (O'Meara et al. 2006; Revell 2008). We focus on the first of these approaches, which we call "phylogenetic prediction."

When using the term phylogenetic prediction, we are specifically referring to predicting trait values on a tip of a tree, in contrast to the occasional use of the phrase "predicting phylogeny" to infer phylogenetic relationships. Just as it is useful to reconstruct traits at internal nodes on a phylogeny, predictions for values of traits on the tips of the tree are valuable for evolutionary research (Garland and Ives 2000; Nunn 2011). For example, predicting values on the tips of the tree can be used to estimate trait values in unmeasured species or for studying species that are too rare and endangered for handling or invasive sampling.

Here, we focus on using phylogenetic prediction to assess whether a species differs from what is expected based on both phylogeny and trait correlations. We might ask, for example, do humans have a later age at first reproduction, relative to other primates and incorporating the body mass scaling of primate life history traits? Prediction would be based on the association between body mass and age at first reproduction and—assuming phylogenetic signal in the traits or residuals from the statistical model (see Chap. 5)—our phylogenetic closeness to other apes. With a prediction in hand, we could then test whether the observed mean age at first reproduction in humans departs from expectations for other primates of our body mass, accounting for broader phylogenetic differences among the species in the sample. It is also possible to account for multiple predictor variables in the prediction, such as diet or predation risk.

In what follows, we consider a general framework for predicting trait values on the tips of the tree in a phylogenetic generalized least squares (PGLS) framework (Garland and Ives 2000; Organ et al. 2007, see Chaps. 5 and 6 for details about PGLS), and we review how this and related approaches have been used in previous comparative research. In the Online Practical Material (http://www.mpcm-evolution.com), we provide new R code and instructions to run the analyses using a Bayesian approach that also performs model selection (e.g., see Chap. 10) and controls for uncertainty in phylogenetic, evolutionary, and statistical parameters. The Online Practical Material also provides further statistical testing of the approach using simulations.

We apply our code—called BayesModelS, for "Bayesian Model Selection"—to two traits in humans in which we predict unique selection pressures relative to other anthropoid primates (monkeys and apes). First, we investigated the intermembral index (IMI). The IMI is calculated as $100 \times$ forelimb length/hindlimb length; it has been used widely in studies of primate morphology because it covaries with categories of locomotor behavior involving vertical clinging and leaping (VCL), quadrupedal, or suspensory locomotion (Napier and Walker 1967; Martin 1990; Napier 1970). The IMI approximates 70 in primate species that exhibit VCL, 70–100 in those with quadrupedal locomotion, and 100–150 in primates that show suspensory locomotion (Martin 1990). Given this strong association, the IMI has been used to reconstruct locomotor behavior in the fossil record (e.g., Napier and Walker 1967; Martin 1990; Jungers 1978). With our highly derived (bipedal) locomotion, humans do not fall into any of these locomotor categories. In addition, our locomotion is associated with long legs and short arms—similar to species showing VCL and low IMI—but we evolved from suspensory species (all apes), which show the largest IMI values. Thus, if you had to bet on a trait as a singularity, the IMI in humans would be a good place to put your money; we test that prediction with our new computer code.

Second, we examine predictors of white blood cell counts in humans, both to re-investigate previous findings with our new methods (Nunn et al. 2000; Nunn 2002) and to test whether humans have exceptionally high numbers of circulating white blood cells. We specifically predicted that humans would have a larger number of neutrophils than predicted for a typical primate because humans have been exposed to a large number of parasites and pathogens through our close

contact with domesticated animals and through agricultural practices (e.g., indirect contact with rodents that raid food stores, more sedentary lives, and formation of vector breeding grounds through irrigation, Barrett et al. 1998). Moreover, cooking likely provided additional energy for humans (Wrangham 2009), potentially increasing investment in immune defenses, which would be reflected by having higher numbers of circulating white blood cells.

21.1 Background: Phylogenetic Prediction

21.1.1 Relevant Literature

One of the first clear descriptions of phylogenetic prediction was provided by Garland and Ives (2000), where they argued for its utility in estimating traits in extinct or unmeasured extant species. They also showed how this approach can be used to identify deviations from allometric relationships, which is similar to our use here—i.e., they provide prediction intervals on a regression and test whether a species falls outside those intervals. Garland and Ives (2000) described how to conduct phylogenetic prediction with either independent contrasts or PGLS and provided approaches for placing confidence intervals on the predictions (see also Garland et al. 1999). In applying the method, Garland and Ives (2000) showed that phylogenetic information provides better predictions of trait values in unmeasured species, specifically by shifting the predicted interval to reflect phylogenetic propinquity to other species in the dataset and narrowing the interval compared to "generic" predictions that lack phylogenetic placement of the unmeasured species.

Organ et al. (2007) used phylogenetic prediction to investigate the evolution of genome size in birds, with a focus on extinct species. It is thought that smaller genomes reduce metabolic costs and are under selection for decreased size in birds due to the energetic expenditure of flight (Hughes and Hughes 1995). Organ et al. (2007) predicted genome size based on the size of osteocytes (bone cells). To do this, they used a Bayesian phylogenetic regression analysis implemented in the program BayesTraits (Pagel and Meade 2007). First, they confirmed an association between osteocyte size and genome size in living vertebrates. Next, they generated posterior probability distributions of genome size in 31 extinct dinosaurs, controlling for uncertainty in phylogeny and statistical parameters with their Bayesian approach. Remarkably, for all but one of the extinct theropods within the lineage that gave rise to birds, genome sizes fell within the range of variation found in living birds. Their analyses therefore suggest that the evolution of reduced genome size occurred *before* the evolution of flight. Thus, evolutionary correlations can be used to make phylogenetically informed predictions about traits that do not fossilize, based on both phylogeny and statistical associations with other features that do fossilize (see also Organ and Shedlock 2009).

This approach has also been used to investigate evolutionary singularities in primate evolution, focusing on human dental characteristics and feeding behavior (Organ et al. 2011). Cooking food is a key behavior that has influenced many aspects of human evolution and is unique to humans (Wrangham 2009), but has this behavior influenced quantitative aspects of our feeding behavior and morphology? Again using BayesTraits, analyses revealed that food processing had major impacts on the amount of time humans spend feeding: Our phylogenetic model predicted that we should feed for 48 % of our daily activity budget if we were a typical primate with our body mass, as compared to an extremely low observed value of 4.7 % in humans. In addition, Organ et al. (2011) found that cooking influenced dental morphology, providing a way to pinpoint the timing of transitions to cooking (and other sophisticated food processing) in human evolution. With a Bayesian phylogeny of hominins, they found evidence for a reduction in molar size in *Homo erectus*, with the morphological change suggesting that this hominin species had already adopted significant food processing behavior well before the emergence of modern humans.

Brain evolution is also a topic of great interest in the context of human evolutionary novelty. The human brain is thought to be central to many important aspects of human uniqueness, especially in terms of our cognitive abilities and social learning (Reader and Laland 2002; Deaner et al. 2007), but it remains unclear which parts of the brain are most important for understanding human uniqueness (Sherwood et al. 2012). At a gross level, many quantitative comparative approaches have been taken to assess brain evolution on the lineage leading to *Homo*, which clearly involved rapid, large, and unique changes (Lieberman 2011; Allman and Martin 2000; Sherwood et al. 2008; Martin 1990). In one recent study, for example, Barton and Venditti (2013) investigated whether human frontal lobes are exceptionally large relative to other brain regions in primates. Remarkably, they found no evidence for such effects and also failed to find evidence of elevated evolutionary change in prefrontal white and gray matter (relative to other brain areas) along the human lineage (examining variation in evolutionary rates is the other approach to investigating evolutionary novelty, noted above).

A somewhat different approach to understanding hominin brain evolution was taken by Pagel (2002). In a study of how brain size changed over time across a phylogeny of fossil hominins, he showed how branch-length scaling parameters can be used to investigate the tempo and mode of evolution (e.g., in terms of acceleration of brain size in human evolution). The intercept from his regression model served as a prediction of brain size in the ancestral node, deep in the human lineage. Like the previous example, Pagel's approach is also more closely tied to estimating rates of evolutionary change. A variety of methods have been developed in this regard, with some based on independent contrasts (McPeek 1995) and others based on detecting variation in rates using maximum-likelihood approaches (O'Meara et al. 2006; Revell 2008).

Phylogenetic prediction is not strictly limited to testing for exceptional evolution. Using the method of Garland and Ives (2000), for example, Fagan et al. (2013) predicted measures of maximum population growth rate in poorly known

mammalian species based on life history characteristics. Using cross-validation procedures, they found good agreement between observed and predicted values.

A different set of methods was used to investigate extinction risk in carnivores (Safi and Pettorelli 2010). In this study, the authors modeled threat status in a clade of 192 carnivore species based on phylogeny, geography, and environmental variables, with one goal to assess how well predictions matched empirical threat levels. Using phylogenetic eigenvector regression (Diniz-Filho et al. 1998) and spatial eigenvector filtering (Diniz-Filho and Bini 2005), they found that geography and phylogeny are important predictors of threat status and probably of other biological characteristics. Thus, it is important to include phylogeny and geography in predictive models of extinction risk.

The issue of evolutionary singularities and phylogenetic prediction was discussed in Nunn (2011). The present chapter expands the discussion of phylogenetic prediction and provides code to implement some of the proposals in Nunn (2011).

21.1.2 Implementation of Phylogenetic Prediction

The first well-described implementation of phylogenetic prediction was provided by Garland and Ives (2000). Using the software PDTREE and two traits, the authors described how to re-root the tree such that the target species and its sister occur at the base of the tree. The user then makes predictions based on the value of X in the target species and the branch length connecting it to the rest of the tree. The authors also provide equations for implementing phylogenetic prediction in a PGLS framework.

As noted above, BayesTraits has also been used to predict trait values based on a regression model (Organ et al. 2007, 2011). The program works well for multiple predictor variables, it can analyze results across a block of trees to account for phylogenetic uncertainty, and it is possible to estimate three different parameters that scale the phylogeny to reflect the degree of phylogenetic signal (see Chap. 5 and below). One issue, however, is that BayesTraits is incompletely documented (prediction is not mentioned in the manual, yet it can be accomplished), and it lacks the flexibility of running analyses within a statistical package that allows programming, such as R. With R code, for example, users can more flexibly automate and adjust the analyses—e.g., combine them easily with other data transformation and statistical procedures of interest—and it is easier to test the assumptions of the methods. An additional issue is that BayesTraits does not provide an easy-to-use model selection procedure that is integrated with parameter estimation. While one could run all possible models and select among them based on likelihood or Bayesian approaches, that would be extremely time-consuming. Based on these issues, we re-implemented and extended much of the functionality of BayesTraits in R.

21.1.3 General Approach

We focus on a case in which the evolutionary singularity occurs on a terminal branch and thus is autapomorphic. The species in which the hypothesized exceptional character occurs is termed the "target species" (Fig. 21.1). We are interested in how some quantitative (continuously varying) trait differs between the target species and the rest of the species in the clade. We call this trait the "target trait." Before proceeding, it is important to note that we are focusing on the situation where we have an a priori expectation for a singularity in some trait(s) in a particular lineage, rather than searching post hoc for exceptional differences predicting a single trait's value for all species in a clade in a one-by-one procedure through all the species.

Our phylogenetic prediction approach involves three steps. First, we build a regression model to describe how a set of independent variables predicts a response variable, where the response variable is the target trait and the target species is *not* included in the analysis. Second, we use the resulting regression model to predict values of the target trait in the target species, based on measured values of the independent variables for the target species and its phylogenetic position relative to the other species in the dataset. Finally, we compare the predicted value of the target trait to the actual value in the target species. The larger this difference, the more "exceptional" the target trait is in the target species relative to other species in the clade, given the statistical model and phylogeny.

We wish to emphasize that phylogenetic prediction is not simply ancestral state reconstruction, as might be used for inferring the states of interior nodes on the tree. Our approach makes use of a phylogeny, an evolutionary model, and output from a regression analysis, whereas trait reconstruction uses only a phylogeny and an evolutionary model (or the assumption of parsimony). The regression analysis incorporates phylogenetic information, typically through a variance–covariance matrix in PGLS (see Chap. 5). Importantly, we exclude the target species when estimating parameters of the PGLS model. The prediction is based on this model, the variance–covariance matrix that includes the target species and predictor variables for the target species. If the statistical model has no predictors (or if the predictors fail to account for variation in the regression model), the approach is similar to standard ancestral state reconstruction, with only phylogeny, an esti-mated intercept, and the underlying evolutionary model providing predictions for the target trait in the target species (i.e., the estimate is made based on deviations from the estimated intercept). If there is no phylogenetic signal in the model, predictions are based solely on the statistical model, as might occur for predictions from a standard least-squares regression.

The method that we develop is based on PGLS, which provides a flexible approach to studying correlated evolution (Pagel 1997, 1999; Garland and Ives 2000; Martins and Hansen 1997; Grafen 1989; Rohlf 2001, Chap. 5). In addition, the statistics estimated within the PGLS framework can be used to reveal other important questions about trait evolution, including the degree of phylogenetic
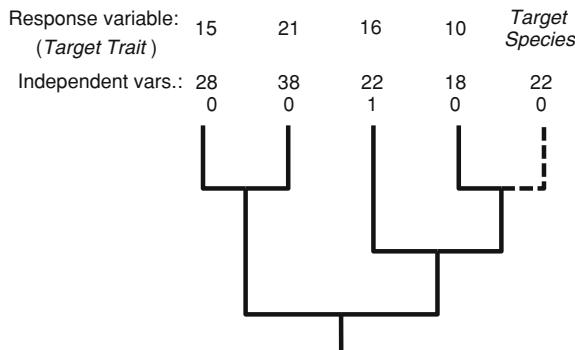
Fig. 21.1 Implementation of phylogenetic targeting. This phylogeny shows a simple case in which we are making a prediction for the "target species" based on two independent variables (the second of which is a binary variable, e.g., nocturnal vs. diurnal activity in primates). We typically have data on the target species, but it is omitted here to emphasize that the regression model is estimated without information on the target species. Then, with independent variables for the target species, predictions for the response are made that incorporate phylogeny and the degree of phylogenetic signal in the residuals of the model (Revell 2010). Finally, the actual value of the response in the target species is compared statistically to the predicted value

signal (i.e., the parameter λ, Pagel 1999; Freckleton et al. 2002). Specifically, we assume that the target trait is the response variable in a PGLS with one or more predictor variables, of the form:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_i X_i + \varepsilon$$

In this model, X and Y are trait values, α is the intercept, β is the regression slope, and ε is the error term. Phylogenetic relatedness is incorporated into the error term with a phylogenetic variance–covariance matrix, which is derived from the phylogenetic topology and branch lengths and scaled to reflect the degree of phylogenetic signal (Pagel and Meade 2007; Freckleton et al. 2002; Nunn 2011).

As noted above, a variety of statistical approaches can be used to assess whether the predicted value of the target trait differs from expectations in the target species. One could simply examine the magnitude of the difference, although this would not provide a way to judge the "significance" of the difference. It is also possible to use likelihood-based methods, for example through a likelihood ratio test. From such a test, a p-value can be obtained, which gives the probability of obtaining a difference as extreme or more extreme, assuming the null hypothesis of no difference is true. Finally, the user can (and should) calculate prediction intervals on the predicted value, again incorporating phylogeny (for equations, see Garland et al. 1999; Garland and Ives 2000).

Here, we use Bayesian framework to assess departures from predictions. Specifically, our method generates a posterior probability distribution for the target trait in a target species—such as a predicted value for body mass in *Homo sapiens*. While many approaches could be taken, including the frequentist approaches just

described, Bayesian approaches are particularly appropriate for phylogenetic prediction. First, they provide a quantitative measure of the degree of difference, measured as the proportion of posterior predictions that are more or less extreme than observed. In addition, the Bayesian framework takes into account uncertainty in other parameters—including uncertainty associated with phylogenetic topology, branch lengths, and estimation of regression parameters and phylogenetic signal (Pagel and Lutzoni 2002). Finally, model selection procedures can be easily implemented within the Bayesian framework and by doing so effectively consider uncertainty in the model selection procedure itself. For Bayesian approaches see Chap. 10.

21.1.4 BayesModelS: R Scripts for Model Fitting, Model Selection, and Prediction

With these needs in mind, we developed new R scripts to conduct phylogenetic prediction in a Bayesian framework. BayesModelS uses a Bayesian Markov Chain Monte Carlo (MCMC) approach to obtain posterior probabilities of regression coefficients and phylogenetic scaling parameters (λ and κ) across a set of trees. The parameter λ is generally viewed as a measure of phylogenetic signal (Freckleton et al. 2002). It scales the off-diagonal elements of the variance–covariance matrix by λ, with $\lambda = 0$ equivalent to no phylogenetic signal because the internal branches collapse to zero length, resulting in a star phylogeny (Felsenstein 1985). The parameter κ raises branch lengths to the exponent κ (Pagel 1997, 2002). Thus, when $\kappa = 0$, all branches are set to be equal, which is consistent with a speciational model of evolution in which change occurs during the process of speciation (Garland et al. 1993), assuming no extinction on the tree. However, we view these parameters as ways to scale the branches to best meet the assumptions of the regression model (especially homoskedasticity), rather than being informative of the underlying evolutionary process.

Our script implements model selection procedures using Bayesian approaches by updating a vector indicating whether variables are included (1) or excluded (0) in the model at steps in the Markov chain and estimating coefficients for those that are included. The specific details of this procedure are provided in the Appendix. When a "missing" species is identified, it is considered to be a target species. Another function in our script then generates a posterior probability distribution of predicted trait values for that species (provided predictor variables are given for the target species).

BayesModelS takes two files as input: one that contains one or more phylogenies with branch lengths (in "phylo" format, Paradis et al. 2004) and the other serving as a data file. The data file should include headers that indicate variable names horizontally along the top, with species names in the first column that

correspond to species in the phylogeny file. The code compares species names in the data and tree files to identify mismatches or missing species and reports those discrepancies to the user. The user specifies the statistical model to be evaluated based on column names, including the possibility to fix a variable in the model so that it is always included (i.e., permanently set to "1" in the vector associated with model selection). The user can also estimate scaling parameters λ and κ or allow the MCMC procedure to select among them. Although it is possible to include λ and κ in the model selection procedure, we recommend also running analyses in which only one of these scaling parameters is estimated, to assess whether similar estimates are obtained. In addition, λ and κ can be fixed to particular values (most commonly 0 or 1). The user also defines a burnin period and sampling (thinning) rate.

Output includes a summary of the data and phylogeny used, a detailed log of parameters and likelihoods in the sampled MCMC iterations, and graphical output of parameter estimates. For phylogenetic prediction, the method produces graphical and quantitative output for assessing whether a target species departs from predictions. Likelihood of the data for sampled models is recorded to provide a way to assess whether burnin was reached (i.e., whether the MCMC chain reached a stable distribution that can be sampled) and whether the thinning rate is sufficient (i.e., with low correlation between adjacent saved states in the chain).

More details are provided in the Appendix and the Electronic Online Material that accompany this book, including mathematical details on how the approach was implemented, simulation tests to assess the performance of the method, R code, data, and instructions for running the functions. The worked examples that follow also demonstrate the utility of the output for assessing performance of the MCMC analysis (e.g., ensuring adequate burnin and using the post-burnin samples).

21.2 Empirical Applications: How Do Humans Differ from Other Primates?

Bipedal Locomotion and the Intermembral Index. We investigated whether the intermembral index (IMI) in humans differs from what one would predict in a primate of our body mass. As described above, we predict that humans have a small IMI compared to other primates because bipedal locomotion has resulted in longer legs relative to arms and thus a lower IMI (see above and Nunn 2011). We tested this prediction using data on 117 primate species and a block of 100 trees sampled from a Bayesian posterior probability distribution provided in version 3 of *10kTrees* (Arnold et al. 2010).

BayesModelS can take multiple predictor variables. Here, as a first illustration of the method, we used a simple model with only body mass as a predictor of IMI. We set the burnin to 100 iterations and sampled the chain every 100 iterations thereafter for 200,000 iterations, resulting in 2,000 samples for the posterior

distribution of all parameters and the prediction for humans. We estimated λ as a measure of phylogenetic signal by setting the argument varSelection (i.e., `var-Selection = ''lambda''`). Analyses were repeated three times to ensure that they stabilized on the same parameter space, and we visually assessed whether likelihoods reached a steady state with adequate sampling. Data were $\log_{10}$-transformed prior to analysis. Humans were identified as "missing" in the analysis through the argument "missingList" in BayesModelS (i.e., `missing-List = c(''Homo_sapiens'')`). The species supplied to missingList are excluded from the first step of estimating parameters in the model (along with all species for which a predictor or response variables are missing in the proposed regression model). Thus, predictions that come from the model are not biased by extreme values in the target species.

When estimating parameters of the statistical model, the chain quickly reached stationarity (Fig. 21.2), showing a plateau with variation typical of an MCMC sample around the plateau (some large moves downward, but disproportionately sampling parameter space that made the data more likely). Body mass was selected to be included in the model in 1991 of the 2000 saved iterations (99.6 %, where model "inclusion" is indicated by zero–one codes and an estimated nonzero regression coefficient in the program output). Moreover, the regression coefficient was consistently positive when it was included in the model (Fig. 21.3, mean = 0.052). We also found good evidence for phylogenetic signal, with the estimate of λ close to 1 (mean = 0.99, Fig. 21.4). For comparison, we also ran the analysis in the caper (Orme et al. 2011) package of *R* using a consensus tree. The estimated regression coefficient was 0.050, and the maximum-likelihood estimate of λ was 1. In this case, BayesModelS provided good agreement to PGLS approaches for this dataset. We recommend that users also cross-check their results with other programs until the performance of our scripts has been more thoroughly tested and check the Online Practical Materials regularly or contact the authors for the latest version of the code.

We then used the model to predict the IMI in humans and compared the predictions to the empirical estimate for humans (72, i.e., 1.86 after $\log_{10}$-transformation). This was achieved by predicting the value for humans based on the saved parameters and the sample of trees from *10kTrees* (Arnold et al. 2010). More specifically, for each of the 2,000 posterior samples of the regression coefficients and other parameters (including λ), we predicted the value in humans, ending up with 2,000 predicted values (Fig. 21.5). The mean predicted value was 2.05, with a 95 % credible interval of 1.99–2.10. Thus, the human value of 1.86—shown as a dashed line in Fig. 21.5a—clearly falls outside this range, as predicted. In addition, the model clearly narrowed in on a narrow portion of the distribution of the IMI across species (Fig. 21.5b). Data and code to run this example is provided in the Online Practical Material.

Primate White Blood Cell Counts. As a second example, we examined primate white blood cell (WBC) counts, focusing on the most common of the WBCs, neutrophils, which are involved in innate immune responses. For reasons outlined
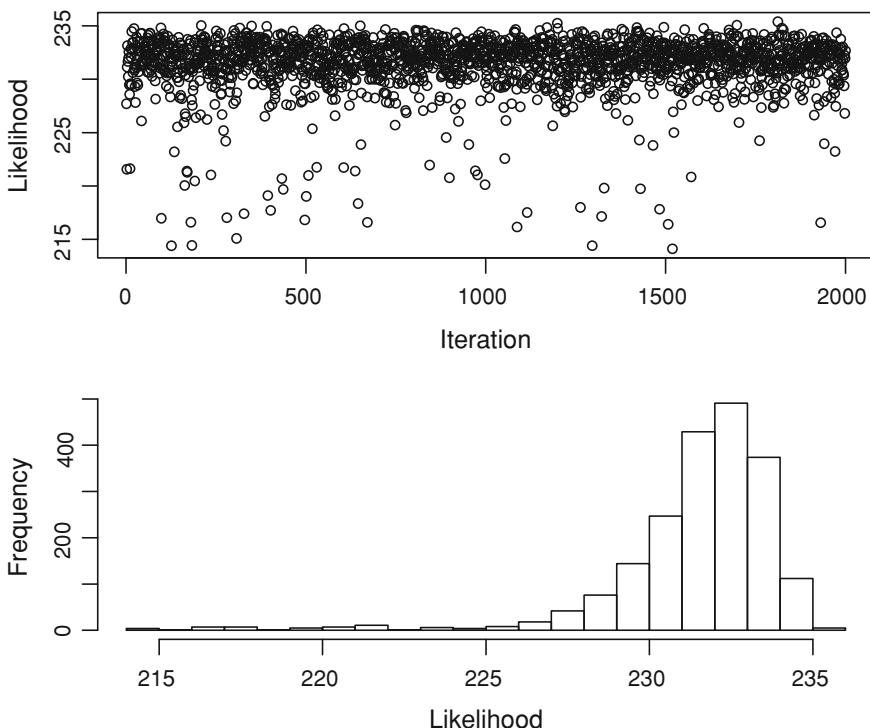
Fig. 21.2 Likelihood of the data based on model parameters across the MCMC run. When run effectively, the MCMC analysis will sample parameter space in proportion to the likelihood of the data (*top panel*). Input also included 100 dated phylogenies (Arnold et al. 2010), one of which was randomly selected at each iteration. The resulting distribution of likelihoods shows that the MCMC chain preferentially samples the models that make the data most likely (*bottom panel*)

above, we predicted that humans would be evolutionary outliers in white blood cell counts. In addition to testing an interesting hypothesis in human evolution, we chose this example because two of the variables are discrete traits with three ordinal levels: Female promiscuity was coded as a three-level variable (monogamous, usually one male but not always, and typically more than one male, from van Schaik et al. 1999); terrestriality was also coded on a three-part scale reflecting typically arboreal, typically terrestrial in wooded environments, and typically terrestrial in an open environment, where the last category is associated with the greatest terrestrial substrate use (from Nunn and van Schaik 2002). We also used the model to show how λ or κ models of trait evolution can be selected using MCMC during estimation of regression coefficients. All data were $\log_{10}$-transformed prior to analysis.

The underlying regression model re-examines previous findings (Nunn et al. 2000; Nunn 2002). Following these studies, we predicted that neutrophil counts increase with promiscuity (to reduce risk of STD transmission), body mass (to reduce
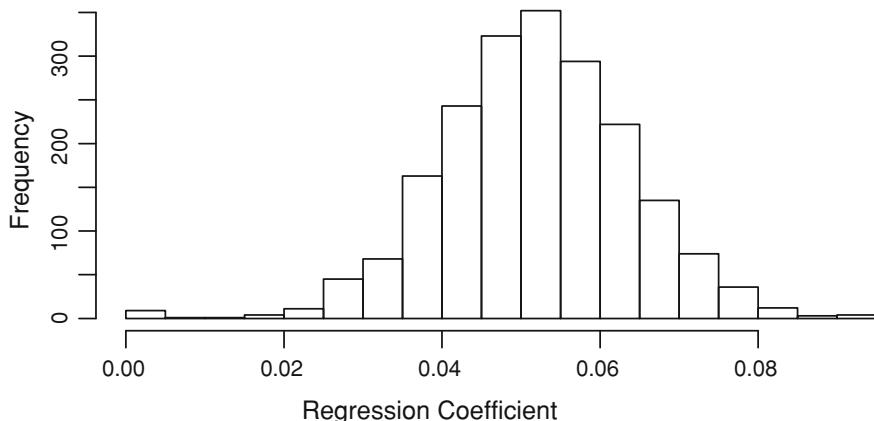
Fig. 21.3 Regression coefficients for the intermembral index (IMI) analysis. Histogram provides the posterior probability of the regression coefficients relating the IMI to body mass from the MCMC run
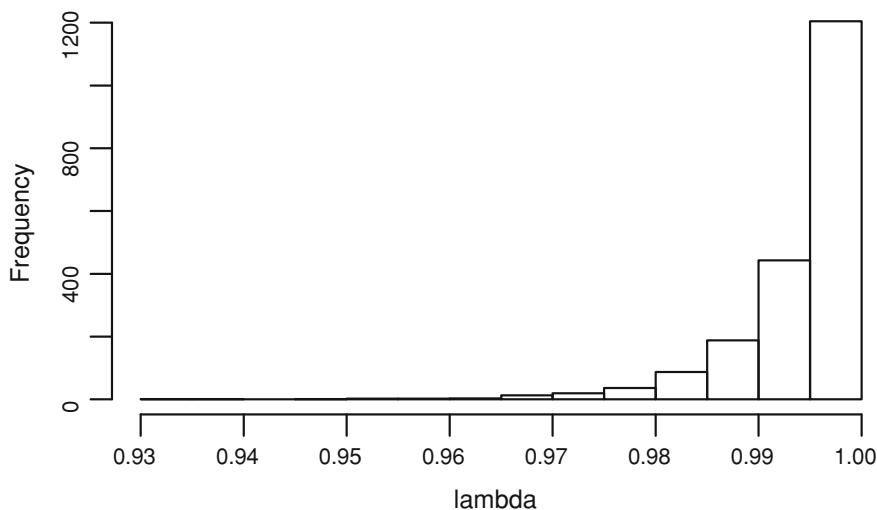
Fig. 21.4 Branch-length scaling parameter in the intermembral index analysis. Posterior probability distribution of the λ parameter (κ was not used)

risk of dietary transmission, with larger animals eating more resources), terrestrial substrate use (to reduce risk of fecally transmitted parasites on the ground), and group size (to reduce risk of social transmission in larger groups). We also included body mass because body mass covaries with some of our variables and larger-bodied primates have been found to have higher WBC counts in previous comparative analyses (Nunn et al. 2009; Cooper et al. 2012). Previous research found support for

(a)

(b)
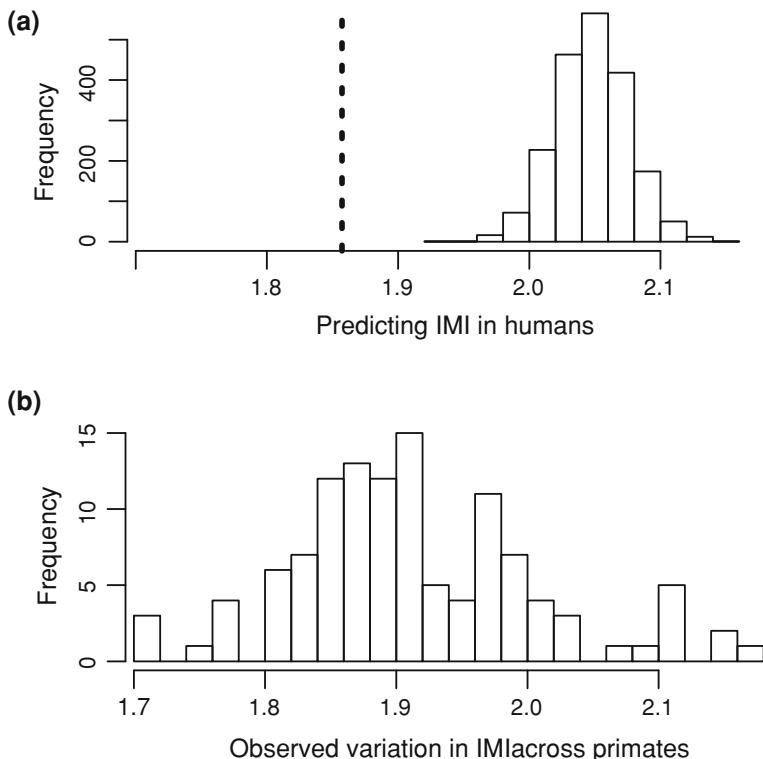
Fig. 21.5 Predicted IMI in humans. **a** *Panel A* shows the posterior probability distribution of predicted IMI for humans, with the empirical (observed) value of the IMI for humans shown as a dashed line well below the predicted values (all values $\log_{10}$-transformed). **b** *Panel B* shows the observed distribution of the IMI in primates as a whole. While many primates have a low IMI, they tend to be small-bodied, and humans come from ancestors with high IMI (suspensory apes). Thus, the observed value for humans is relatively low, based on our body mass and phylogenetic position

the effects of promiscuity, but not group size or substrate use (Nunn et al. 2000; Nunn 2002).

Thus, we estimated the regression coefficients for the following full model, while also using model selection procedures to determine which variables should be included in the model:

$$\text{Neutrophils} \sim \text{Group Size} + \text{Body Mass} + \text{Promiscuity}\{0, 1, 2\}$$
$$+ \text{Substrate}\{0, 1, 2\}$$

Values in the curly brackets indicate the levels of the "promiscuity" and "substrate" use variables, with increasing values indicating increasing promiscuity or terrestrial substrate use, respectively. We used the same MCMC settings that were

used in the IMI analysis, giving 2,000 posterior distribution samples of the regression coefficients and other parameters and thus 2,000 predictions for humans. For estimating λ versus κ, we set `varSelection = ''random.''`

The analysis revealed that mating behavior (promiscuity) was entered into the model in all sampled runs, with effects as demonstrated previously (Nunn 2002). That is, more promiscuous species have higher neutrophil counts (Fig. 21.6). As with previous analyses, we failed to find support for effects of terrestrial substrate use (only 1.2 % of models) and group size (only 6.6 % of models) as predictors of neutrophil counts. Body mass was also rarely included in the model (only 10.2 % of the models). Interestingly, we found greater support for using the κ-transformation than the λ-transformation, with 87.6 % of the posterior sample favoring κ and a relatively low estimate of κ (mean = 0.092, see Fig. 21.7). This suggests that many comparative analyses that focus only on estimating λ may be doing a suboptimal job of transforming phylogeny for comparative analyses. We also ran the analysis treating the factor variables as if they were continuously varying. This produced weaker results for the promiscuity codes, with the variable included in the model in only 65.3 % of the models (but positive in 99.6 % of the samples in which it was entered).

To predict neutrophil counts in humans, we used the coefficients from the "factor" models (i.e., discrete codes for promiscuity and substrate use), assuming that humans are monogamous. Given that the promiscuity variable is discrete and neither of the quantitative variables (body mass or group size) was consistently entered into the models, we expected a wide 95 % credible interval. That is what we found, with a distribution (0.45–0.97) that is nearly as wide as the observed distribution of values in primates (Fig. 21.8; compare this to the substantially narrower distribution of predicted IMI relative to other primates in Fig. 21.5). The observed value for humans falls within this wide distribution, resulting in no evidence for human uniqueness in terms of neutrophil counts. Without quantitative predictors, however, the model gives only a weak prediction for this test, making it hard to draw firm conclusions. Computer code, phylogenies, and data for these analyses are provided in the Online Practical Material.

21.2.1 Future Directions and Conclusions

Evolutionary novelties pose a serious challenge for the comparative method. Biologists would like to study these traits in broad comparative perspective, but how can this be achieved without falling into the trap of adaptive storytelling? How can we investigate evolutionary singularities—or cases of "exceptional evolution" for quantitative traits—in a statistically rigorous way? We propose that phylogenetic prediction offers a valuable solution for this challenge, especially when combined with other methods, such as comparing rates of evolutionary change in different lineages (O'Meara et al. 2006; Revell et al. 2008). In particular, the underlying statistical model is based on widely accepted approaches to
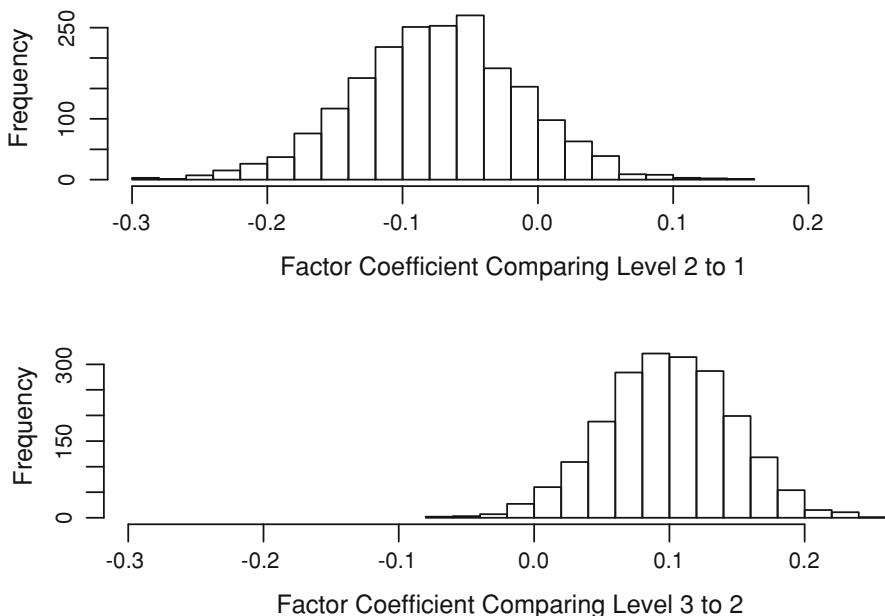
Fig. 21.6 Regression coefficients for the promiscuity codes in analyses of white blood cell counts. BayesModelS generated two dummy variables that compare intermediately promiscuous primates to those that are monogamous (*top*) and another that compare intermediately promiscuous primates to those that are highly promiscuous (*bottom*). The negative coefficient for the *top plot* indicates that less promiscuous primates have lower white blood cell counts, while the positive coefficient for the *bottom plot* indicates that more promiscuous primates have higher white blood cell counts. The latter comparison is especially strong, with more of the distribution of coefficients being positive

investigating adaptive evolution using phylogenetic comparative methods. When this model is applied to a single lineage in a phylogenetic context, it gives fresh insights into whether the general pattern of adaptive evolution is also explanatory in the "target species" of interest.

The importance of phylogenetic prediction is underappreciated in studies of the evolutionary process, especially when considering whether a particular species departs from the overall evolutionary pattern in a group of organisms. The approach has yet to be widely used, due in part to lack of good implementation in *R*, which is becoming established as the standard for comparative analyses. In this chapter, we aimed to overcome these limitations through new *R* scripts (Bayes-ModelS), original analyses, and supplementary datasets that enable others to run our analyses. We focused especially on the context of human evolution. However, we expect that many other biological systems provide similar examples of evolutionary singularities for which this perspective—and versions of our code—would be useful. It is worth noting that our BayesModelS code can also be run without the prediction component to provide Bayesian PGLS with model selection.
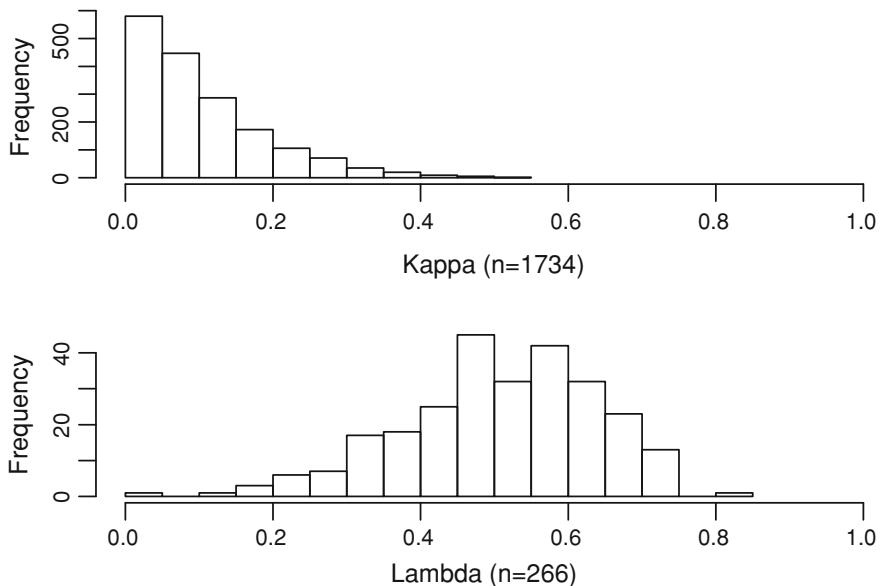
Fig. 21.7 Posterior probability distribution of the branch-length scaling parameters. The λ or κ parameters were selected as part of the MCMC analysis (the κ model was preferred, indicated by the larger sample in the posterior distribution for κ than for λ)

In terms of future directions, it would be desirable to further assess the statistical performance of BayesModelS to detect differences (some simulations are provided in the Appendix). By simulating evolutionary change under known conditions and varying the number of species, it is possible to investigate Type I error rates (when simulation along a lineage uses the same model as in other species) and Type II error rates (i.e., statistical power, when higher rates of evolution and/or directional evolution occur on the target species' lineage, corresponding to greater evolutionary change). By using a particular phylogeny in the simulation—e.g., primates if the question involves humans as the target species—one could estimate statistical properties in a specific biological context of a hypothesized singularity in a particular lineage. It will also be important to investigate whether predictive capability declines when more extreme values of predictor variables are used, especially if those extreme values are more likely to result in singularities through nonlinear or threshold effects or if they involve extrapolation beyond available data.

As noted throughout, one can also investigate exceptional evolution in terms of variable rates of evolution: We expect elevated rates of evolution in the target trait on the lineage leading to the target species. In a previous application of this general approach to study human feeding time, for example, Organ et al. (2011) showed that the rate of evolution was substantially elevated in the lineage leading to humans. Treating branch length as a measure of evolutionary rate, the branch
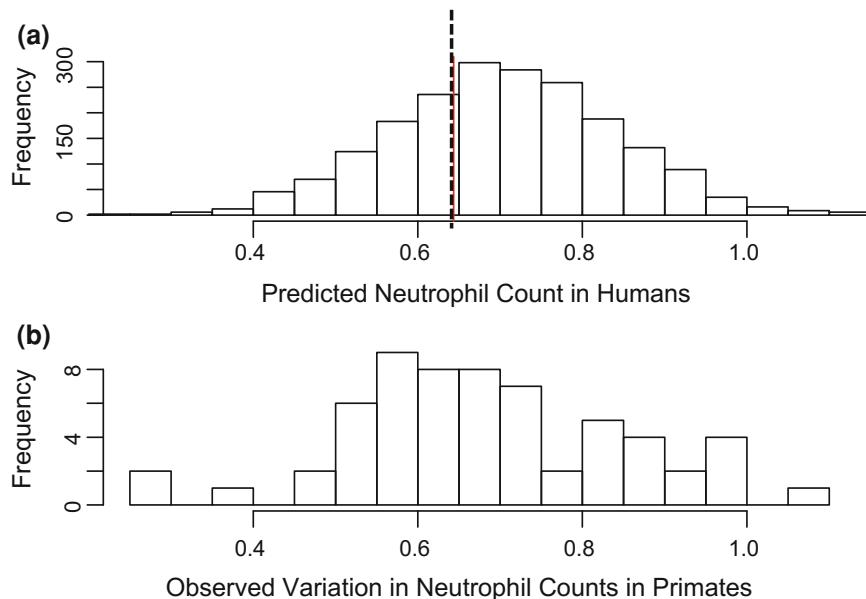
Fig. 21.8 Human neutrophil count predicted. Posterior probability distribution of predicted number of neutrophils in humans. **a** *Panel A* shows the posterior probability distribution of predicted neutrophils for humans, with the empirical (observed) value for humans shown as a vertical *dashed line* (all values $\log_{10}$-transformed). The observed value falls within the posterior probability distribution, which is exceptionally wide (see main text). **b** *Panel B* shows the observed distribution of neutrophil counts in primates as a whole

leading to humans would need to be *50 times longer* to accommodate the large reduction in molar size in early *Homo* (under a Brownian motion model and based on changes in body mass). In another example using evolutionary rate variation, Nunn (2011) applied a method (McPeek 1995) based on independent contrasts to study the IMI. As expected, this analysis revealed that change on the branch leading to modern humans is significantly elevated, as compared to other contrasts among the apes. Thus, in addition to computer code provided here for phylogenetic prediction, it would be valuable to develop user-friendly code to implement a wide range of methods, such as McPeek's (1995) method, or to study variable rates of evolution in the context of singularities using existing code, such as Brownie (O'Meara et al. 2006) and related code in the phytools package in R (Revell 2011).

Appendix:
Phylogenetic Prediction for Extant and Extinct Species

1. *Mathematical Description of Method*

Consider the following regression model for n different species:

$$y_i = \alpha_0 + \theta_1 x_{i1} + \cdots + \theta_m x_{im} + \in_i \qquad (21.1)$$

In the above model, y_i is the response variable for the ith species and $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{im})$ are covariates associated with the ith species. The error terms for all species $\in = (\in_1, \in_2, \ldots, \in_n)$ follow multivariate normal distribution:

$$\in \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{V\sigma^2})$$

In this equation, $\mathbf{V}$ is the covariance matrix structure and σ^2 is the standard deviation. Ordinary linear regression usually assumes the errors are independent, identical, and normally distributed, such that the covariance matrix has the same value along the diagonal of $\mathbf{V}$ with off-diagonal set to zero. For biological data, however, different species will exhibit similarity because of common ancestry, which leads to positive values on the off-diagonals. Moreover, the diagonal of V may show heterogeneity if root-to-tip distances vary, as might be the case if fossils are included or when the branch lengths are based on molecular change rather than absolute dates. As noted above, it is possible to select scaling parameters that transform the branch lengths to better model the evolution of traits on a given tree topology. The parameter λ scales internal branches (off-diagonal elements of V) between 0 and 1; when $\lambda = 0$, this corresponds to no phylogenetic structure, i.e., a star phylogeny. The parameter κ raises all branches to the power κ. Thus, when $\kappa = 0$, this corresponds to a phylogeny with equal branch lengths, as might occur when speciational change takes place.

Hence, the covariance structure $\mathbf{V}$ can be crucial to comparative analyses of species values, and scaling parameters provide important insights into the evolutionary process and degree of phylogenetic signal in the data.

The objective is to select the optimal model with respect to different covariates and variance structures. Two variance structures, λ and κ, are considered as scaling parameters. We aim to select covariates as well as the variance structure that best characterizes trait evolution. Meanwhile, precise estimation of different parameters (regression coefficients, λ, κ) is also required. Given that only λ and κ are considered, we can rewrite the distribution of $\in$ as follows:

$$\in \sim \mathcal{N}(\boldsymbol{0}, (\boldsymbol{I_V \sigma_\lambda^2} + (\boldsymbol{1} - \boldsymbol{I_V})\boldsymbol{\sigma_\kappa^2})\boldsymbol{\Sigma}(\boldsymbol{T}, \boldsymbol{I_V}, \boldsymbol{\lambda}, \boldsymbol{\kappa})) \qquad (21.2)$$

In this equation, I_V indicates the selection of variance structure. I_V will be equal to 1 for estimating λ and 0 for estimating κ. The parameters $\sigma_\lambda^2, \sigma_\kappa^2$ are standard deviations for λ and κ. Covariance matrix $\Sigma(T, I_V, \lambda, \kappa)$ is a function of the evolutionary tree T, indicator I_V, λ, and κ. Henceforth, we will use notation Σ to replace $\Sigma(T, I_V, \lambda, \kappa)$.

Using a Bayesian framework, the parameters are treated as random variables and their distribution is investigated. In order to select models, three types of parameters are included in the above model:

1. Parameters for tree selection T. Here, we would use a large number of trees to represent uncertainty in the phylogeny that describes evolutionary relationships among the species. A posterior distribution of M trees $\{T_1, \ldots, T_M\}$ will be used and treated as a uniform distribution (although a single tree can also be used).
2. Parameters for variable selection $\Theta_1 = (\gamma, \boldsymbol{\beta})$. This includes the indicator variable $\gamma = (\gamma_1, \ldots, \gamma_m)$, which indicates whether a variable is included in the model. Moreover, effect size $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)$ for each covariate is also included. The regression coefficient $\theta_i = \gamma_i \times \beta_i, i = 1, 2, \ldots, m$.
3. Parameters for variance selection $\Theta_2 = (I_V, \lambda, \kappa, \sigma_\lambda^2, \sigma_\kappa^2)$

Let Y and X be matrices of the response variable and m explanatory variables for n species, respectively, as given below:

$$Y = (y_1, y_2, \ldots, y_n)^T, \quad X = \begin{pmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{pmatrix}$$

Then, the joint posterior distribution for all parameters will be

$$f(\gamma, \boldsymbol{\beta}, I_V, \lambda, \kappa, \sigma_\lambda^2, \sigma_\kappa^2, T | Y, X) \propto p(\gamma, \boldsymbol{\beta}, I_V, \lambda, \kappa, \sigma_\lambda^2, \sigma_\kappa^2, T) f(Y | \gamma, \boldsymbol{\beta}, I_V, \lambda, \kappa, \sigma_\lambda^2, \sigma_\kappa^2, T, X) \quad (21.3)$$

In the above equation,

1. $p(\gamma, \boldsymbol{\beta}, I_V, \lambda, \kappa, \sigma_\lambda^2, \sigma_\kappa^2, T)$ is the prior distribution for all parameters. We assume the priors of tree selection, variable selection parameters and variance selection parameters are independent, i.e., $p(\gamma, \boldsymbol{\beta}, I_V, \lambda, \kappa, \sigma_\lambda^2, \sigma_\kappa^2, T) = p(T)p(\gamma, \boldsymbol{\beta})$ $p(I_V, \lambda, \kappa, \sigma_\lambda^2, \sigma_\kappa^2, T)$. We also assume prior to variable selection, $p(\gamma, \boldsymbol{\beta})$, to satisfy $p(\gamma, \boldsymbol{\beta}) = p(\gamma)p(\boldsymbol{\beta}|\gamma)$. $p(\gamma)$ follows a non-informative prior, and for each i, $\beta_i | \gamma_i = (1 - \gamma_i)N(\hat{\mu}, S) + \gamma_i N(0, \tau^2)$, while $\hat{\mu}, S, \tau^2$ are predefined parameters.
2. $f(Y | \gamma, \boldsymbol{\beta}, I_V, \lambda, \kappa, \sigma_\lambda^2, \sigma_\kappa^2, T, X)$ is the probability density function:

$$f(Y | \gamma, \boldsymbol{\beta}, I_V, \lambda, \kappa, \sigma_\lambda^2, \sigma_\kappa^2, T, X) = I_V N(Y | X\theta, \sigma_\lambda^2 \Sigma) + (1 - I_V)N(Y | X\theta, \sigma_\kappa^2 \Sigma)$$

Equation (21.3) is difficult to analyze. However, $\boldsymbol{\beta}$ can be integrated out, which significantly simplifies the calculation. Consequently, we only need to consider the posterior distribution $f(\gamma, I_V, \lambda, \kappa, \sigma_\lambda^2, \sigma_\kappa^2, T | Y, X)$. After the posterior distribution is obtained, $f(\boldsymbol{\beta} | \gamma, I_V, \lambda, \kappa, \sigma_\lambda^2, \sigma_\kappa^2, T, Y, X)$ follows multivariate normal distribution.

Let $X(\gamma)$ be columns of X with $\gamma_j = 1$ and $\Sigma' = \Sigma'(T, I_V, \lambda, \kappa, \sigma^2) = (\frac{1}{\sigma^2} X(\gamma)^T \Sigma^{-1} X(\gamma) + \frac{1}{\tau^2} I)$, then the posterior distribution can be simplified to:

$$f(\gamma, I_V, \lambda, \kappa, \sigma_\lambda^2, \sigma_\kappa^2, T | Y, X) = p(\gamma) p(I_V, \lambda, \kappa, \sigma_\lambda^2, \sigma_\kappa^2) p(T)$$

$$\times \left(I_V \frac{\det(\Sigma')^{\frac{1}{2}}}{(\sigma_\lambda^2)^{(\sum \gamma)} \det(\Sigma)^{\frac{1}{2}}} \exp\left(-\frac{1}{2\sigma_\lambda^2} A_1\right) + (1 - I_V) \frac{\det(\Sigma')^{\frac{1}{2}}}{(\sigma_\kappa^2)^{(\sum \gamma)} \det(\Sigma)^{\frac{1}{2}}} \exp\left(-\frac{1}{2\sigma_\kappa^2} A_2\right) \right)$$

$$(21.4)$$

where $A_1 = Y' \Sigma^{-1} Y - \frac{1}{\sigma_\lambda^2} (y' \Sigma^{-1} X(\gamma) \Sigma' X(\gamma)' \Sigma^{-1} Y)$, and $A_2 = Y' \Sigma^{-1} Y - \frac{1}{\sigma_\kappa^2} (y' \Sigma^{-1} X(\gamma) \Sigma' X(\gamma)' \Sigma^{-1} Y)$.

The posterior distribution from Eq. (21.4) is difficult to obtain; hence, we generate posterior samples using MCMC (Liu 2003) to select the optimal model. Gibbs sampling will be used to get the posterior samples. Gibbs sampling is an algorithm that can generate a sequence of samples from the joint probability distribution of two or more random variables. In each iteration of Gibbs sampling, we use the following procedure to obtain posterior samples:

1. Simulate T_k from $\{T_1, \ldots, T_M\}$;
2. Simulate γ from $f(\gamma | I_V, \lambda, \kappa, \sigma_\lambda^2, \sigma_\kappa^2, T, Y, X)$;
3. Simulate I_V from $f(I_V | \gamma, \lambda, \kappa, \sigma_\lambda^2, \sigma_\kappa^2, T, Y, X)$;
4. Simulate $\lambda, \kappa, \sigma_\lambda^2, \sigma_\kappa^2$ from $f(\lambda, \kappa, \sigma_\lambda^2, \sigma_\kappa^2 | \gamma, I_V, T, Y, X)$;
5. Simulate $\boldsymbol{\beta}$ from $f(\boldsymbol{\beta} | \gamma, \lambda, \kappa, \sigma_\lambda^2, \sigma_\kappa^2, T, Y, X)$.

Since γ and I_V in Step 1 follow a Bernoulli distribution, the posterior sample can be directly obtained. In Step 4, λ, κ can be obtained by the Metropolis–Hasting Algorithm (Hastings 1970) and $\sigma_\lambda^2, \sigma_\kappa^2$ from the inverse gamma distribution. $\boldsymbol{\beta}$ in Step 5 follows a multivariate normal distribution.

After N posterior samples, $\left\{\Theta_1^{(1)}, \Theta_1^{(2)}, \ldots, \Theta_1^{(N)}\right\}$, $\left\{\Theta_2^{(1)}, \Theta_2^{(2)}, \ldots, \Theta_2^{(N)}\right\}$ and $\{T^{(1)}, T^{(2)}, \ldots, T^{(N)}\}$ have been obtained, we are interested in which model should be selected, which can be achieved via different criteria and goals:

1. *Model with the highest posterior probability*

Let $P(M_i | X, Y), i = 1, 2 \ldots, 2^{m+1}$ be the posterior probability of ith candidate models. $P(M_i | X, Y)$ is given by the percentage of model i in the posterior samples.

The one with the highest posterior probability can be selected as the optimal model:

$$M_* = \text{argmax}_i P(M_i|X, Y)$$

2. *Inclusion probability for variables (model selection)*

The inclusion probability of *j*th variable can be defined as $P(\gamma_j = 1|X, Y)$, which is a marginal probability across all posterior samples. This probability can be estimated by $P(\gamma_j = 1|X, Y) = \frac{\sum \gamma_j^{(k)}}{N}$

3. *Probability of the variance structure*

The probability of λ model and κ model, $P(I_V = 1|X, Y)$, can be obtained through $P(I_V = 1|X, Y) = \frac{\sum I_V^{(k)}}{N}$

4. *Inference on regression coefficients*

For a specific candidate model M_i, the inference on parameters for M_i can be directly obtained from posterior samples for M_i. Moreover, estimation of effect size in general for a certain covariates can be obtained through Bayesian model averaging (BMA) (O'Hara and Sillanpaay 2009). For example, β_i, which is effect size for the *i*th covariates, can be estimated as posterior mean:

$$\hat{\beta}_i = \sum_k P(M_k|X, Y)E_{M_k}(\beta_i)$$

where $E_{M_k}(\beta_i)$ is the average of posterior sample β_i for M_k model. The above estimator is actually the general mean of posterior sample for $\beta_i^{(k)}$. So we can use

$$\text{Var}\left(\hat{\beta}_i\right) = \text{Var}(\beta_i^{(k)})$$

as the estimator for the variance of $\hat{\beta}_i$.

Model Checking

Bayesian model checking (Gelman 2004) can be used to check whether the model is consistent with the data. Consider data Y, X and corresponding posterior samples, $\left\{\Theta_1^{(1)}, \Theta_1^{(2)}, \ldots, \Theta_1^{(N)}\right\}, \left\{\Theta_2^{(1)}, \Theta_2^{(2)}, \ldots, \Theta_2^{(N)}\right\}$ and $\{T^{(1)}, T^{(2)}, \ldots, T^{(N)}\}$. Under the assumption of a linear model, we can use each posterior sample to generate one predicted (i.e., "fake") $Y^{(k)}$ given $\{X, T^{(k)}, \Theta_1^{(k)}, \Theta_2^{(k)}\}$ in the following way:

$$Y^{(k)} \sim N\left(X\theta^{(k)}, \left(I_V^{(k)}\left(\sigma_\lambda^2\right)^{(k)} + \left(1 - I_V^{(k)}\right)\left(\sigma_\kappa^2\right)^{(k)}\right)\Sigma^{(k)}\right)$$

(a)
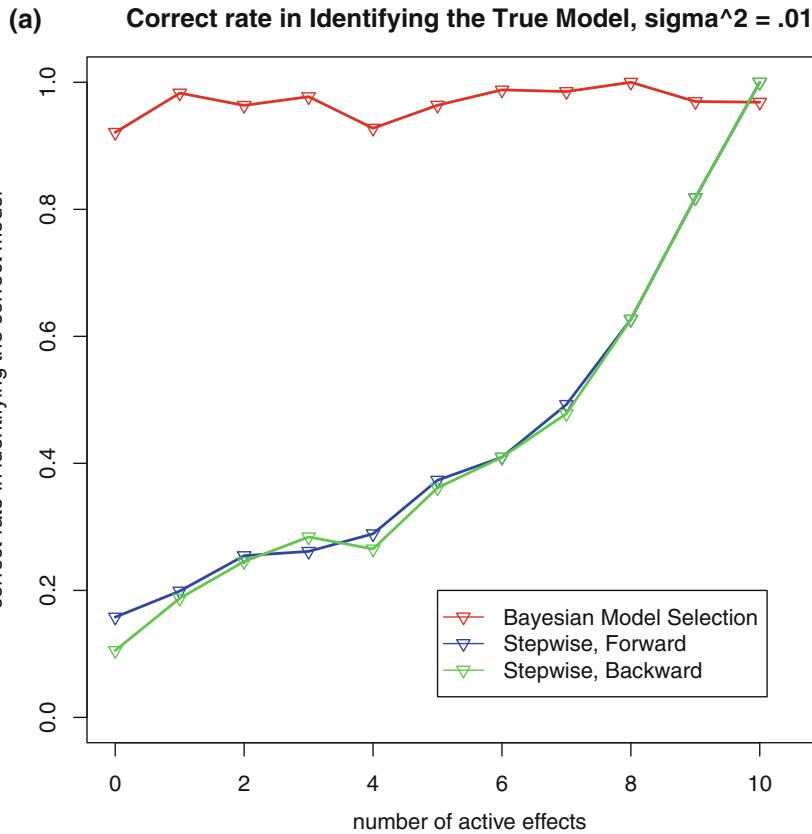
Fig. 21.9 a Percentage of time each method identifies the correct model $(\sigma_\lambda^2, \sigma_\kappa^2 = 0.01)$. **b** Percentage of time each method identifies the correct model $(\sigma_\lambda^2, \sigma_\kappa^2 = 0.02)$. **c** Percentage of time each method identifies the correct model $(\sigma_\lambda^2, \sigma_\kappa^2 = 0.03)$

where $\Sigma^{(k)} = \Sigma(T^{(k)}, I_V^{(k)}, \lambda^{(k)}, \kappa^{(k)})$. So for each posterior sample, one fake $\boldsymbol{Y}^{(k)}$ can be obtained. A predefined function $z^{(k)} = f(\boldsymbol{Y}^{(k)})$ with $\{\boldsymbol{Y}^{(k)}, = 1, 2, \ldots, N\}$ can be obtained and compared to $z_C = f(\boldsymbol{Y})$ obtained through real data. With comparison between $\{z^{(k)}\}$ and z_C, the validity of the model is evaluated.

The logic of model checking is that if the model is valid, then the generated fake $\boldsymbol{Y}$s should be statistically similar to the true observed $\boldsymbol{Y}$. The choice of function f depends on the dataset and model we have used. However, there are several commonly used f functions, e.g., variance $(z^{(k)} = \text{var}(\boldsymbol{Y}^{(k)}))$ and median $(z^{(k)} = \text{median}(\boldsymbol{Y}^{(k)}))$. Each time, we check z_C against the distribution of $\{z^{(k)}\}$ and two-sided p-values will be identified. If p-value is smaller than 0.05, we will conclude that the data are not consistent with the model.

(b)   **Correct rate in Identifying the True Model, sigma^2 = .02**
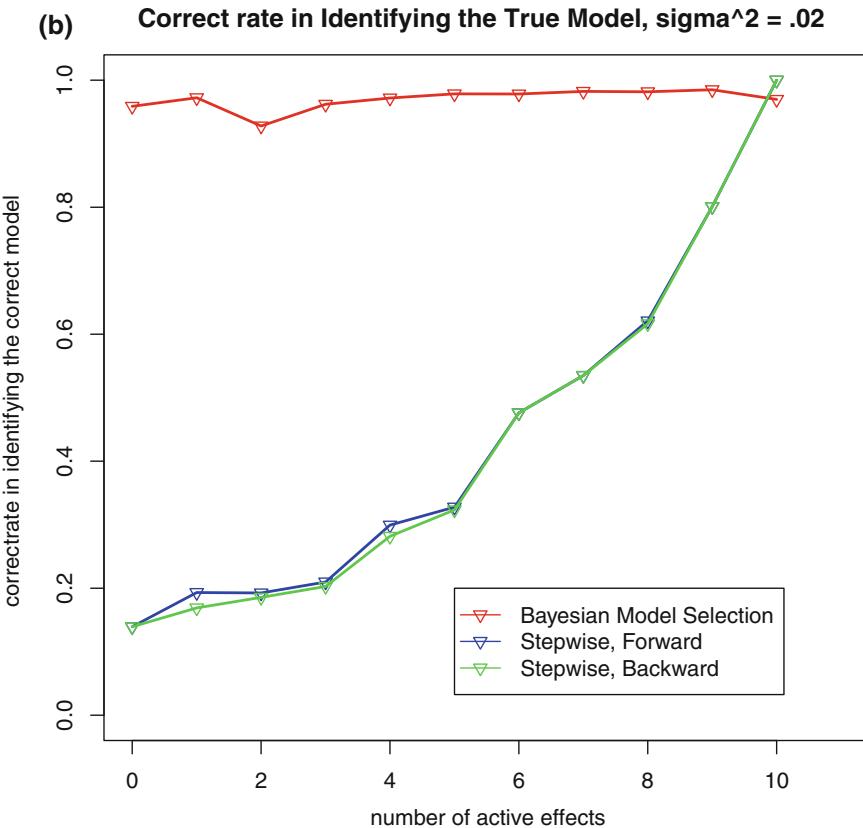
Fig. 21.9 (continued)

Prediction for an unknown response (Gelman 2004) for a species is also available in this Bayesian framework, for example, if one wishes to predict a value for a species that has not yet been studied or to investigate whether a particular species deviates from expectations of the evolutionary model (Organ et al. 2011). Consider species n_1 with known tree T_{new} and explanatory variable X_{new}, but response variable Y_{new} is missing. Assume posterior samples, $\left\{\Theta_1^{(1)}, \Theta_1^{(2)}, \ldots, \Theta_1^{(N)}\right\}$, $\left\{\Theta_2^{(1)}, \Theta_2^{(2)}, \ldots, \Theta_2^{(N)}\right\}$ and $\left\{T^{(1)}, T^{(2)}, \ldots, T^{(N)}\right\}$ are obtained, the joint distribution of $\left(Y^{(k)}, Y_{new}^{(k)}\right)^T$ given $X_{new}, T_{new}, \Theta_1^{(k)}, \Theta_2^{(k)}$ will be:

$$\left(Y^{(k)}, Y_{new}^{(k)}\right) \sim \mathcal{N}\left((X, X_{new})^T \theta^{(k)}, \left(I_V^{(k)}\left(\sigma_\lambda^2\right)^{(k)} + \left(1 - I_V^{(k)}\right)\left(\sigma_\kappa^2\right)^{(k)}\right)\Sigma\left(T^{(k)} \cup T_{new}, I_V^{(k)}, \lambda^{(k)}, \kappa^{(k)}\right)\right)$$
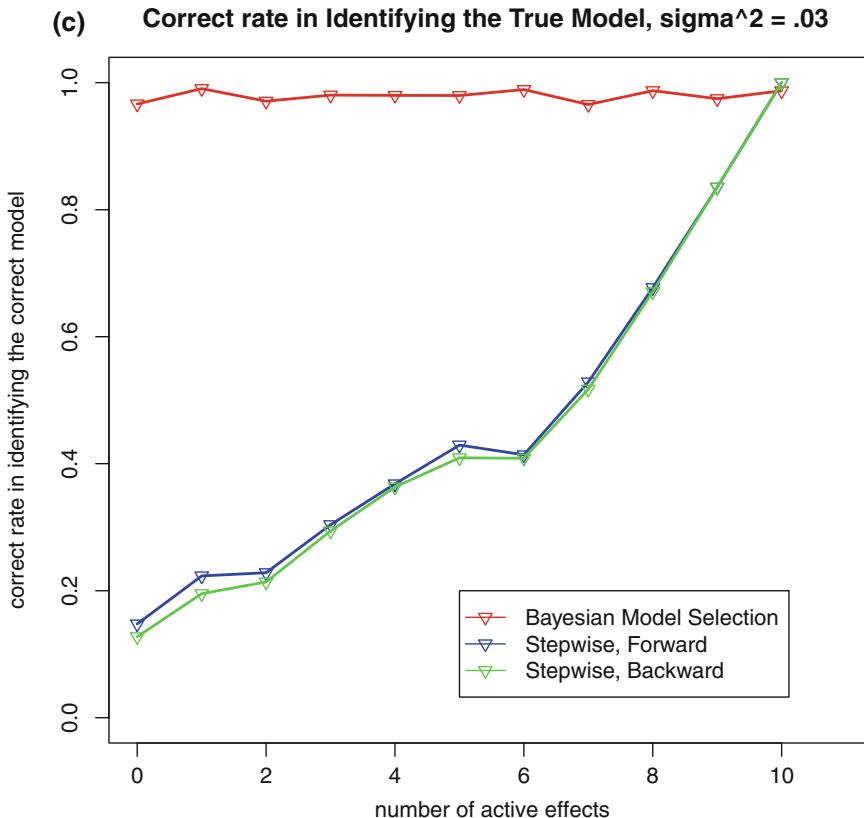
Fig. 21.9 (continued)

for each posterior sample. Let $\Sigma_{\text{new}}^{(k)} = \Sigma\left(T_{\text{new}}, I_V^{(k)}, \lambda^{(k)}, \kappa^{(k)}\right)$, then the covariance matrix for combined tree will satisfy:

$$\Sigma\left(T^{(k)} \cup T_{\text{new}}, I_V^{(k)}, \lambda^{(k)}, \kappa^{(k)}\right) = \begin{pmatrix} \Sigma^{(k)} & \Sigma_{12}^{(k)} \\ \Sigma_{21}^{(k)} & \Sigma_{\text{new}}^{(k)} \end{pmatrix} = \begin{pmatrix} \Sigma^{(k)} & \Sigma_{12}^{(k)} \\ \Sigma_{21}^{(k)} & \Sigma_{22}^{(k)} \end{pmatrix}$$

Since we have already observed y, the distribution of $y_{\text{new}}^{(k)}$ can be obtained through a conditional normal distribution:

$$Y_{\text{new}}^{(k)}|Y \sim N(\overline{\mu^{(k)}}, \overline{\Sigma^{(k)}})$$
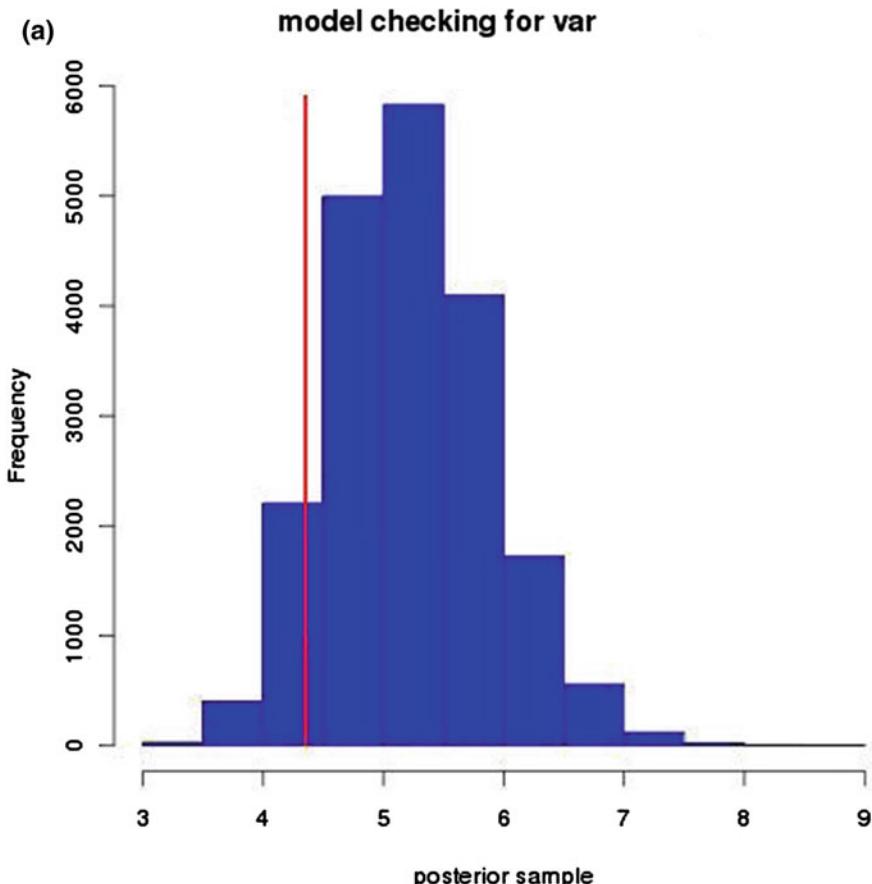
Fig. 21.10 **a** Model checking for variance. **b** Model checking for mean. **c** Model checking for minimum. **d** Model checking for maximum. *Red line* indicates the actual data

while

$$\overline{\mu^{(k)}} = X_{\text{new}} \theta^{(k)} + \Sigma_{21}^{(k)} \left(\Sigma_{11}^{(k)} \right)^{-1} (Y - X\theta^{(k)})$$

$$\overline{\Sigma^{(k)}} = \Sigma_{22}^{(k)} - \Sigma_{21}^{(k)} \left(\Sigma_{22}^{(k)} \right)^{-1} \Sigma_{12}^{(k)}$$

So for each posterior sample, one simulated $Y_{\text{new}}^{(k)}$ can be obtained. Then, we can use the median and variance of predictive draws $\{Y_{\text{new}}^{(k)}, k = 1, 2, \ldots, N\}$ to make predictions for values of the response variable in the new species. If the observed value for the species falls outside of, for example, the 95 % credible interval of
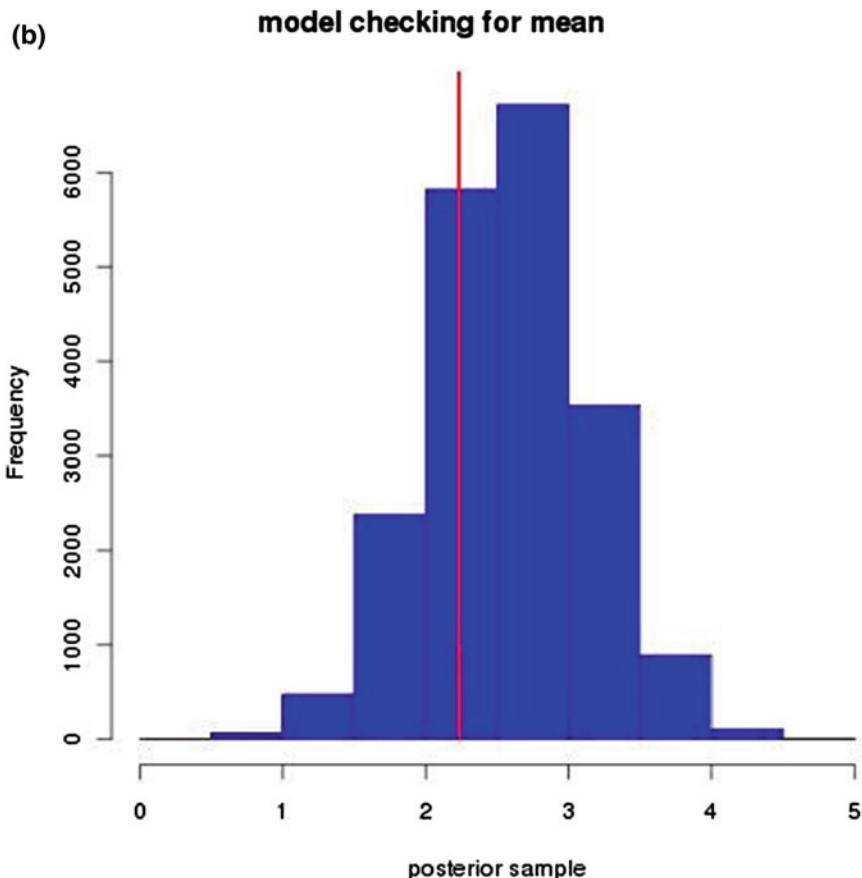
Fig. 21.10 (continued)

predictions, one might infer that an exceptional amount of evolutionary change has occurred.

2. *Simulation Test of Method Implemented in BayesModelS*

We use simulated data to evaluate the performance of BayesModelS, focusing on estimation of parameters (but not prediction). Comparisons between our procedure and stepwise regression were conducted. For each dataset, we simulated predictor variables *X* and response variable *Y* with known associations among the variables on a single phylogeny taken from a posterior distribution of 100 phylogenies for 87 primate species. The variables for each species are independently and identically distributed according to $N(0, 1)$. For BayesModelS, we then ran analyses across 100 trees. For stepwise regression, we used a single tree, which was identical to the tree used to generate the data.
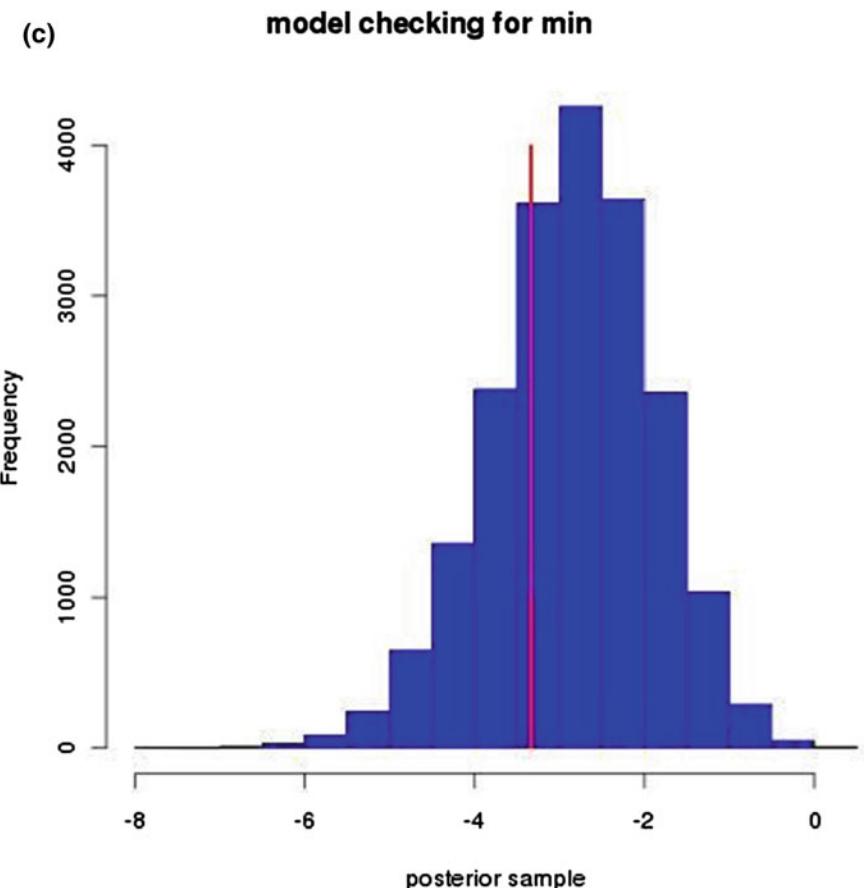
Fig. 21.10 (continued)

Two different sets of simulated data were used. The first dataset is used to check whether Bayesian variable selection correctly identifies the variables to include in the statistical model, as compared to stepwise regression. The inclusion posterior probability of significant and insignificant effects was also evaluated. Consider a regression model with 10 covariates. The coefficients for covariates will be assumed to follow the distribution:

$$\beta_i \sim I_S \times \mathcal{N}\left(\mu, \sigma_1^2\right) + (1 - I_S) \times \mathcal{N}\left(0, \sigma_2^2\right), \forall i$$

while $\mu, \sigma_1^2, \sigma_2^2$ are predefined as $1, 0.1, 0.01$, respectively. I_S is an indicator of whether or not this effect is active (i.e., nonzero). The response variable y can be simulated from Eq. (21.1). Different dataset with $I_V = 1$, $\lambda/\kappa = \text{Unif}[0, 1]$, and $\sigma_\lambda^2, \sigma_\kappa^2 = 0.01, 0.02, 0.03$ will be used.

(d)

model checking for max
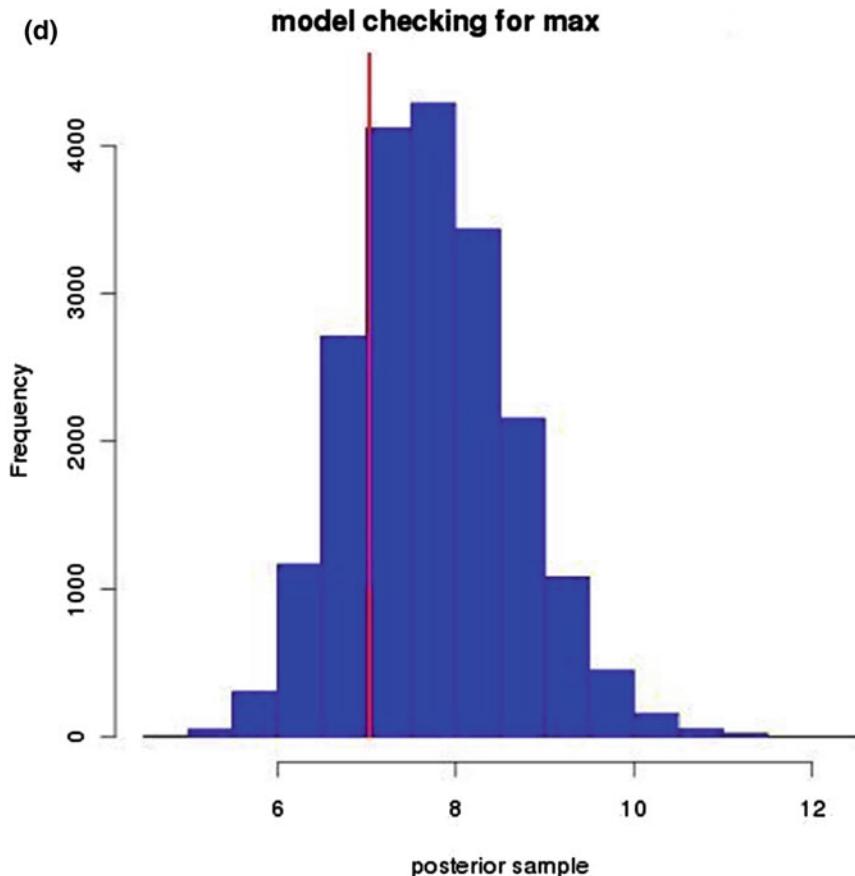
Fig. 21.10 (continued)

In reality, sometimes the true regression coefficients are neither zero nor large (O'Hara and Sillanpaay 2009), as in the previous dataset. The sizes of coefficients can be tapered toward zero. In this part, we consider a regression framework similar to O'Hara and Sillanpaay (2009), with the following regression model:

$$y_i = \alpha + \sum_{j=1}^{m} \theta_j x_{ij} + \in_i$$

For simulations, known values of $\alpha = \log(10)$ and $\sigma_\lambda^2, \sigma_\kappa^2 = 0.01, 0.02, 0.03$ were used. The covariate values, the x_{ij}'s, were simulated independently and drawn from a standard normal distribution, $N(0,1)$. We also assume $m = 21$ and $\in \sim \mathcal{N}(0, \Sigma(T, 1, 0.5, 0.5))$ or $\in \sim \mathcal{N}(0, \Sigma(T, 0, 0.5, 0.5))$, for the models of λ and κ, respectively. The regression coefficients, θ_j, were generated as equal distance
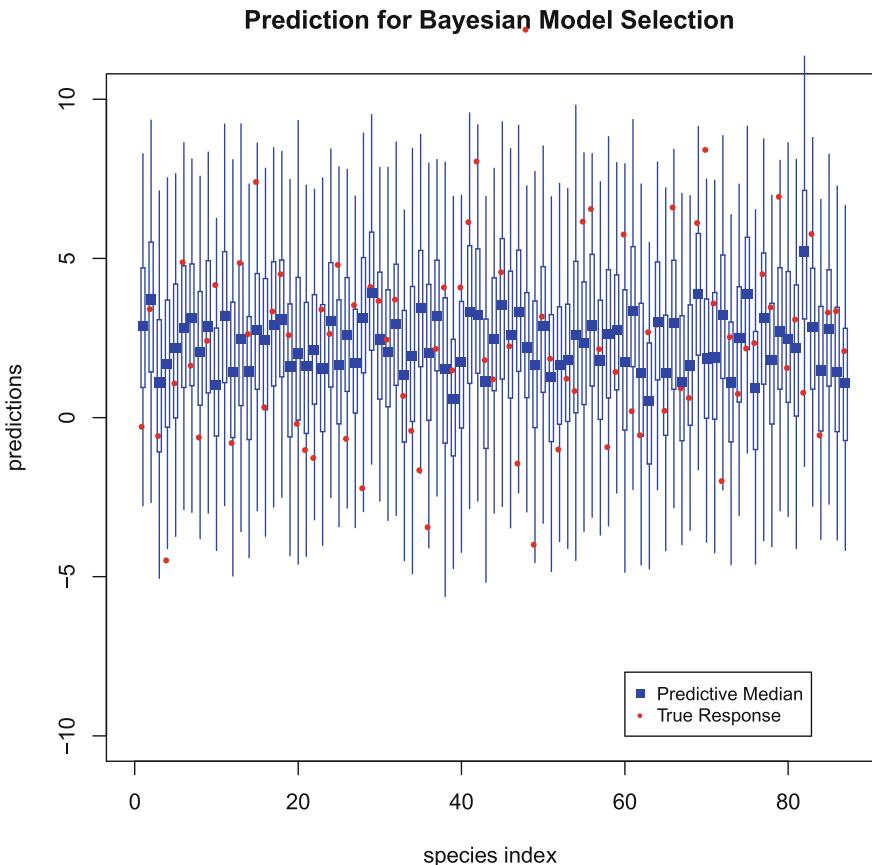
Fig. 21.11 Prediction of 87 species with Bayesian Model Selection. *Blue points* are for median of predictive sample, and *red points* are true response. The *blue squares* are 50 % credible interval, and *blue lines* are 95 % credible interval for predictive samples. Forty of the predictions are outside the 50 % intervals (relative to expectation of 43.5), while 2 of the predictions are outside the 95 % interval (relative to expectation of 4.3)

between $a - bk$ and $a + bk$, while $a = 0, b = 0.05$. Twenty datasets were generated for $k = 1, 2, \ldots, 20$.

We used several performance measures to evaluate BayesModelS. We checked whether BayesModelS can successfully identify the correct model, identified as the model with the highest posterior probability. For the stepwise regression, the optimal model was chosen using both forward and backward stepwise procedures. Repeated simulations were conducted to check the percentage of time the two methods identify the correct model.

Moreover, median of inclusion probability for each covariate in Bayesian Model Selection was also evaluated. This is compared to the percentage of inclusion for each covariate of stepwise regression through repeated simulations.

We do the following for 500 times. Each time, we use a tree to generate data. Then, we use Bayesian method and stepwise regression to estimate the correct model for these data. Since we know the true model, we know whether this is right or wrong for the two methods. We assess the statistical performance of Bayes-ModelS and stepwise regression from this set of results.

The percentage of time each method can identify correct model with the simulated data can be found in the following Fig. 21.9

In more than 90 % of the simulations, the Bayesian Model Selection procedure identified the correct model, regardless of the $\sigma_\lambda^2/\sigma_\kappa^2$ value. It is worth noting that stepwise regression performed well when the number of significant effects is high. When the number of significant effects is low, stepwise regression performs poorly due to a high Type I error rate (Mundry and Nunn 2009).

Next, model checking and prediction with Bayesian Model Selection were conducted. One fixed sample from Data 2 was simulated with $I_V = 1$, $\lambda = 0.5, \sigma_\lambda^2 = 0.1, k = 20$. We used four functions to check validity of the model, mean, variance, median, and range. The checking result can be found in Fig. 21.10. We find that the model is consistent with the data, which is not surprising since the data are generated from line model.

Finally, we used BayesModelS to predict unknown species in a simulation context. Each time, one species was identified as "missing" and then the predict() function was used to predict this species based on the remaining 86 species. The predictive sample was compared to the true response, as shown in Fig. 21.11. We can find that the true response of most species is within 95 % confidence interval of prediction, which means Bayesian Model Selection can effectively make prediction on unknown species.

References

Allman JM, Martin B (2000) Evolving brains. Scientific American Library, Nueva York

Arnold C, Matthews LJ, Nunn CL (2010) The *10kTrees* website: a new online resource for primate phylogeny. Evol Anthropol 19:114–118

Barrett R, Kuzawa CW, McDade T, Armelagos GJ (1998) Emerging and re-emerging infectious diseases: the third epidemiologic transition. Annu Rev Anthropol 27:247–271

Barton RA (1996) Neocortex size and behavioural ecology in primates. Proc R Soc Lond (Biol) 263:173–177

Barton RA, Venditti C (2013) Human frontal lobes are not relatively large. PNAS 110:9001–9006

Cooper N, Kamilar JM, Nunn CL (2012) Longevity and parasite species richness in mammals. PLoS One

Deaner RO, Isler K, Burkart J, van Schaik C (2007) Overall brain size, and not encephalization quotient, best predicts cognitive ability across non-human primates. Brain Behav Evol 70:115–124

Deaner RO, Nunn CL, van Schaik CP (2000) Comparative tests of primate cognition: different scaling methods produce different results. Brain Behav Evol 55:44–52

Diniz-Filho JAF, De Sant'ana CER, Bini LM (1998) An eigenvector method for estimating phylogenetic inertia. Evolution 52:1247–1262

Diniz-Filho JAF, Bini LM (2005) Modelling geographical patterns in species richness using eigenvector-based spatial filters. Global Ecol Biogeogr 14:177–185

Dunbar RIM (1993) Coevolution of neocortical size, group size and language in humans. Behav Brain Sci 16:681–735

Fagan WF, Pearson YE, Larsen EA, Lynch HJ, Turner JB, Staver H, Noble AE, Bewick S, Goldberg EE (2013) Phylogenetic prediction of the maximum per capita rate of population growth. Proc R Soc Lond (Biol) 280:20130523

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

Freckleton RP, Harvey PH, Pagel M (2002) Phylogenetic analysis and comparative data: a test and review of evidence. Am Nat 160:712–726

Garland T, Bennett AF, Rezende EL (2005) Phylogenetic approaches in comparative physiology. J Exp Biol 208:3015–3035

Garland T, Dickerman AW, Janis CM, Jones JA (1993) Phylogenetic analysis of covariance by computer simulation. Syst Biol 42:265–292

Garland T, Ives AR (2000) Using the past to predict the present: confidence intervals for regression equations in phylogenetic comparative methods. Am Nat 155:346–364

Garland T, Midford PE, Ives AR (1999) An introduction to phylogenetically based statistical methods, with a new method for confidence intervals on ancestral values. Am Zool 39:374–388

Gelman A (2004) Bayesian Data Analysis. Chapman & Hall/CRC, London/Boca Raton

Grafen A (1989) The phylogenetic regression. Philos Trans R Soc Lond (Biol) 326:119–157

Harvey PH, Pagel MD (1991) The comparative method in evolutionary biology., Oxford Series in Ecology and EvolutionOxford University Press, Oxford

Hastings WK (1970) Monte Carlo sampling methods using Markov Chains and their applications. Biometrika 57(1):97–109

Hughes AL, Hughes MK (1995) Small genomes for better flyers. Nature 377:391. doi:10.1038/377391a0

Jungers WL (1978) Functional significance of skeletal allometry in megaladapis in comparison to living prosimians. Am J Phys Anthropol 49:303–314

Kappeler PM, Silk JB (eds) (2009) Mind the gap: tracing the origins of human universals. Springer, Berlin

Lieberman D (2011) The evolution of the human head. Belknap Press, Cambridge

Liu J (2003) Monte Carlo strategies in scientific computing. Springer, Berlin

Maddison WP, Midford PE, Otto SP (2007) Estimating a binary character's effect on speciation and extinction. Syst Biol 56:701–710

Martin R (2002) Primatology as an essential basis for biological anthropology. Evol Anthropol 11:3–6

Martin RD (1990) Primate origins and evolution. Chapman and Hall, London

Martins EP (1994) Estimating the rate of phenotypic evolution from comparative data. Am Nat 144:193–209

Martins EP, Hansen TF (1997) Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. Am Nat 149:646–667

McPeek MA (1995) Testing hypotheses about evolutionary change on single branches of a phylogeny using evolutionary contrasts. Am Nat 145:686–703

Mundry R, Nunn CL (2009) Stepwise model fitting and statistical inference: turning noise into signal pollution. Am Nat 173:119–123

Napier JR (1970) The roots of mankind. Smithsonian Institution Press, Washington

Napier JR, Walker AC (1967) Vertical clinging and leaping—a newly recognized category of locomotor behaviour of primates. Folia Primatol 6:204–219

Nee S (2006) Birth-death models in macroevolution. Ann Rev Ecol Evol S 37:1–17

Nunn CL (2002) A comparative study of leukocyte counts and disease risk in primates. Evolution 56:177–190

Nunn CL (2011) The comparative approach in evolutionary anthropology and biology. University of Chicago Press, Chicago

Nunn CL, Gittleman JL, Antonovics J (2000) Promiscuity and the primate immune system. Science 290:1168–1170

Nunn CL, Lindenfors P, Pursall ER, Rolff J (2009) On sexual dimorphism in immune function. Philos Trans Roy Soc B Biol Sci 364:61–69. doi:10.1098/Rstb.2008.0148

Nunn CL, van Schaik CP (2002) Reconstructing the behavioral ecology of extinct primates. In: Plavcan JM, Kay RF, Jungers WL, Schaik CPv (eds) Reconstructing behavior in the fossil record. Kluwer Academic/Plenum, New York, pp 159–216

O'Hara RB, Sillanpaay MJ (2009) A review of bayesian variable selection methods: what, how and which. Bayesian Anal 4(1):85–118

O'Meara BC, Ane C, Sanderson MJ, Wainwright PC (2006) Testing for different rates of continuous trait evolution using likelihood. Evolution 60:922–933

Organ CL, Nunn CL, Machanda Z, Wrangham RW (2011) Phylogenetic rate shifts in feeding time during the evolution of *Homo*. Proc Natl Acad Sci USA 108:14555–14559

Organ CL, Shedlock AM (2009) Palaeogenomics of pterosaurs and the evolution of small genome size in flying vertebrates. Biol Lett 5:47–50

Organ CL, Shedlock AM, Meade A, Pagel M, Edwards SV (2007) Origin of avian genome size and structure in non-avian dinosaurs. Nature 446:180–184

Orme D, Freckleton R, Thomas G, Petzoldt T, Fritz S, Isaac N (2011) Caper: comparative analyses of phylogenetics and evolution in R. http://R-Forge.R-project.org/projects/caper/

Pagel M (1997) Inferring evolutionary processes from phylogenies. Zool Scr 26:331–348

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

Pagel M (2002) Modelling the evolution of continuously varying characters on phylogenetic trees: the case of hominid cranial capacity. In: MacLeod N, Forey PL (eds) Morphology, shape and phylogeny. Taylor and Francis, London, pp 269–286

Pagel M, Lutzoni F (2002) Accounting for phylogenetic uncertainty in comparative studies of evolution and adaptation. In: Lässig M, Valleriani A (eds) Biological evolution and statistical physics. Springer, Berlin, pp 148–161

Pagel M, Meade A (2007) Bayes traits (http://www.evolution.rdg.ac.uk). 1.0 edn., Reading, UK

Pagel MD (1994) The adaptationist wager. In: Eggleton P, Vane-Wright RI (eds) Phylogenetics and Ecology. Academic, London, pp 29–51

Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290

Reader SM, Laland KN (2002) Social intelligence, innovation, and enhanced brain size in primates. PNAS 99:4436–4441

Revell L (2010) Phylogenetic signal and linear regression on species data. Methods Ecol Evol 1:319–329

Revell LJ (2008) On the analysis of evolutionary change along single branches in a phylogeny. Am Nat 172:140–147

Revell LJ (2011) Phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol Evol

Rodseth L, Wrangham RW, Harrigan AM, Smuts BB, Dare R, Fox R, King B, Lee P, Foley R, Muller J, Otterbein K, Strier K, Turke P, Wolpoff M (1991) The human community as a primate society. Curr Anthropol 32:221–254

Rohlf FJ (2001) Comparative methods for the analysis of continuous variables: geometric interpretations. Evolution 55:2143–2160

Safi K, Pettorelli N (2010) Phylogenetic, spatial and environmental components of extinction risk in carnivores. Global Ecol Biogeogr 19:352–362

Sherwood CC, Bauernfeind AL, Bianchi S, Raghanti MA, Hof PR (2012) Human brain evolution writ large and small. In: Hofman M, Falk D (eds) Evolution of the primate brain: from neuron to behavior, vol 195. Elsevier, Amsterdam, pp 237–254

Sherwood CC, Subiaul F, Zawidzki TW (2008) A natural history of the human mind: tracing evolutionary changes in brain and cognition. J Anat 212:426–454

Tennie C, Call J, Tomasello M (2009) Ratcheting up the ratchet: on the evolution of cumulative culture. Philos Trans R Soc Lond (Biol) Biol Sci 364:2405–2415

Tooby J, DeVore I (1987) The reconstruction of hominid behavioral evolution through strategic modeling. In: Kinzey WG (ed) The evolution of human behavior: primate models. State University of New York Press, Albany, pp 183–237

van Schaik CP, van Noordwijk MA, Nunn CL (1999) Sex and social evolution in primates. In: Lee PC (ed) Comparative primate socioecology. Cambridge University Press, Cambridge, pp 204–240

Wrangham RW (2009) Catching fire: how cooking made us human. Basic Books, New York

Chapter 22
Preparing Paleontological Datasets for Phylogenetic Comparative Methods

David W. Bapst

Abstract The fossil record holds considerable promise for furthering our understanding of macroevolutionary patterns, particularly allowing us to analyze hypotheses which cannot be tested with phylogenies of extant taxa alone. However, although there is a growing number of paleontological studies that use phylogenetic comparative methods to address questions of trait evolution, there is little documentation on obtaining the timescaled phylogenies of fossil taxa required for such analyses. This chapter is an attempt to introduce interested readers to the issues involved with that process, including the uncertainties and biases involved with fossil data, which some might inadvertently overlook. In addition, I illustrate how the fossil records of different groups can be very different in terms of the datasets available, including the issues of that data, and stress that there is no 'one size fits all' solution. Instead, for several hypothetical examples, I recommend several approaches that explicitly consider potential uncertainties, unavailable data, and biasing factors.

22.1 Introduction

Phylogeny-based analyses of evolution have proven critical to testing for macroevolutionary processes and measuring the tempo of diversification and trait evolution. Increasingly, biologists working on extant taxa (i.e., 'neontologists') have available a large number of timescaled, ultrametric phylogenies or, if not, tools are readily available to obtain such a phylogeny, given the necessary data. However, information from the fossil record can reveal patterns that cannot even be predicted from datasets consisting of extant taxa alone (Finarelli and

D. W. Bapst (✉)
South Dakota School of Mines and Technology, Rapid City, SD, USA
e-mail: dwbapst@gmail.com

Flynn 2006; Losos 2010; Slater et al. 2012; Slater 2013). Working with fossil data is the only avenue available to paleontologists who wish to understand evolution in extinct groups. Unfortunately, the typical procedures for obtaining and preparing a phylogeny for analyses often do not apply to fossil datasets. This chapter is intended to fill this paleontological gap in the primers that are presently available for preparing data for general comparative analyses and to introduce users to the basic concepts and necessary tools for evaluating the fossil record.

This chapter is structured in a series of discussions with the reader. First, I address the current state of analyses in paleontology that depend on phylogenetic datasets, particularly comparative studies of trait evolution, as it is currently an active area of interest and the theme of this book. Secondly, using several hypothetical workers with similar research questions, I illustrate how the information and data obtainable for different groups of organisms can vary drastically, due simply to differences in morphological completeness, incompleteness of the associated geologic record, taphonomy and the historical contingencies of the paleontological approaches used in studying that group. Finally, I discuss several approaches that take into consideration uncertainties and biases of fossil data, rather than ignoring such issues.

22.1.1 Questions for Phylogeny-Based Analyses in Paleobiology

For decades, paleobiologists have used phylogenetic and non-phylogenetic datasets to understand a myriad of topics in macroevolution. To cover every sort of analysis that a paleobiologist might attempt with a phylogenetic dataset would be impossible. In line with the theme of this volume, this chapter will specifically focus on preparing fossil datasets for analyzing trait evolution, which are often approached with a family of analyses referred to as phylogenetic comparative methods. Borrowing phylogenetic comparative methods from the neontological literature to study trait evolution has recently become a popular paleobiological pursuit (for some recent examples: Friedman 2009; Hunt and Carrano 2010; Hopkins 2011, 2013; Benson et al. 2012; Fusco et al. 2012; Sallan and Friedman 2012; Benson and Choiniere 2013; Raia et al. 2013; Zanno and Makovicky 2013). The majority of the methods used were generally developed with only ultrametric molecular phylogenies in mind and so it is often not apparent what issues need to be considered before applying these methods to fossil data. Some typical questions addressed with these analyses of trait evolution are 'Was there a directional bias over time in the evolution of a trait, such as is postulated for body size under Cope's Rule?' 'Were rates of trait change relatively faster or slower in a particular group or interval?' or 'What is the relationship among traits or between a trait and an environmental predictor?' Analyses for trait evolution are often quite similar to analyses used in model-based studies of phylogeography, which have also recently

been applied to fossil data (Gates et al. 2012), and such biogeographic analyses usually require similar types of phylogenetic datasets.

Why would we want to analyze trait evolution with information from the fossil record, given that including this information can introduce issues not encountered with extant lineages? Other than the obvious reason that the clade of interest might be extinct and only known from the fossil record, we also know that paleontological data can reveal patterns that are not apparent in analyses including only living species. Constant, directional trait evolution cannot be reconstructed or detected on ultrametric phylogenies, unless a strong prior is put on the root state (Felsenstein 1988; Cunningham and Oakley 2000). Finarelli and Flynn (2006) showed that a trend of increasing body size could not be reconstructed in canid mammals without fossil data. Slater et al. (2012) used simulations to show that even small amounts of fossil data can assist in identifying patterns of trait evolution, such as trends.

Additionally, comparative analyses that utilize the fossil record can also reveal changes in patterns of evolution through time that would not necessarily be detectable from extant taxa alone, even when active trends are not involved. For example, Slater (2013) used a phylogenetic dataset of extinct and extant mammals to show that Mesozoic mammal appears to have been constrained to evolve only small body sizes. Similarly, Benson and Choinere (2013) used a dataset of extinct Mesozoic theropods to identify a shift toward increased rates of forelimb evolution in early avian dinosaurs. Wagner and Erwin (2006) used a phylogeny and a dataset of discrete ecomorphological characters to test whether Paleozoic gastropod shells were constrained by intrinsic or extrinsic factors. Additionally, paleontological comparative studies can utilize information unavailable for extant species, such as the longevity of morphospecies in the fossil record as a predictor for extinction risk. For example, Hopkins' (2011) phylogenetic study of Cambrian trilobites found that species longevity was correlated with both geographic range and (surprisingly) reduced intraspecific variation in morphology. In a similar study, Roy et al. (2009) used analyses of phylogenetic signal to show that extinction rates among bivalve families were predicted by their relatedness, suggesting that extinction risk is tied to heritable traits.

However, if we have a dataset from the fossil record, why do we need to study trait evolution in phylogenetic context? Most of the questions considered above were first addressed by paleobiologists without the explicit data structures that we refer to as phylogenies. One particular predecessor to phylogenetic approaches was comparing ancestor–descendant pairs to estimate trends (Alroy 1998, 2000). Alroy (1998) used taxonomic structure and the order of appearance within genera to effectively average over the uncertainty in ancestor–descendant relationships within higher taxa. Similar approaches have also been developed to get at rates of trait change in the absence of phylogenetic information (Bapst et al. 2012; Evans et al. 2012). However, these 'non-phylogenetic' approaches still require information about relationships, often using taxonomy as a rough proxy for the phylogenetic hierarchy, assuming that taxonomy is both trustworthy and that earlier taxa are probably ancestors to later appearing taxa.

There are many other popular questions that are not the focus of this chapter; in fact, there are probably enough topics concerning the use of fossil data in phylogeny-based analyses to fill an entire book. However, many of the related analyses are either quite simple and thus do not warrant extended discussion, or the methods available are incomplete and further development of methods is needed.

A major issue for a biological audience would be the selection of fossil calibrations for obtaining a timescaled molecular tree, which can be sensitive to the choice of fossil dating constraints (Benton and Donoghue 2007; Heath 2012; Warnock et al. 2012). This has motivated the development of approaches that rely on additional data from the fossil record for calculating more informed priors (e.g., Nowak et al. 2013). Similar questions in paleobiology simply do not require the methodological preparation discussed in this chapter, such as 'When did a particular clade first diverge from their sister lineage?' or 'How incomplete is the fossil record, given a phylogenetic hypothesis for a group?' Analyzing the misfit between phylogeny and the appearance times of taxa generally only requires a cladogram and set of first appearance times (e.g., Norell and Novacek 1992; Huelsenbeck 1994; Benton and Storrs 1994; Benton and Hitchin 1997; Siddall 1996; Wills 1999; Pol and Norell 2001; Wills et al. 2008; Boyd et al. 2011). Scientists interested in these questions, however, may benefit from the discussion of timescaling approaches in this chapter.

The diversification of lineages has long been a focus of paleobiology, with considerable interest in how information from relationships might be used in our estimates of past diversity and rates of origination (branching) and extinction. Phylogenetic corrections to diversity estimates mainly rely on long established methods (Norell 1992; Smith 1994), although there is debate about the conditions under which phylogenetic information improves taxonomic estimates of diversity and the degree to which potential ancestor–descendant relationships need to be considered (Wagner 1995, 2000; Norell 1996; Lane et al. 2005; Guinot et al. 2012; Bapst 2014). Biologists have developed a wide range of diversification analyses for molecular phylogenies based on tree imbalance (e.g., Mooers and Heard 1997; Chan and Moore 2002) or using branch times from the present (e.g., Nee et al. 1992; Alfaro et al. 2009). None of these methods are entirely appropriate for fossil data due to incomplete sampling, even when using a time-slice approach to create ultrametric phylogenies (Tarver and Donoghue 2011). However, this topic is still in its infancy, with some models recently introduced for paleontological phylogenies that can directly parameterize sampling processes (Stadler 2010; Didier et al. 2012) and new attempts to combine typical taxonomic estimates for obtaining diversification rates from phylogenetic datasets (e.g., Simpson et al. 2011; Ruta et al. 2011). In addition, Ezard et al. (2011) took a very different approach from those mentioned and applied hazard models to a phylogeny of living and extinct planktonic foraminifera to estimate the effect of traits and environment on diversification and extinction.

22.2 The Fossil Record is Different–and Fossil Records are Different

There are numerous reasons why the typical dataset for extant taxa and any given paleontological dataset might require different approaches. Not all of these issues are immediately obvious, but they all require special consideration. A key element of paleontological data is that the available fossil record for each group of organisms can be very different in terms of the type and amount of data available. This means the methods that we can apply to analyze and interpret phylogenetic patterns in one clade may not work *at all* for a different group's fossil record.

To illustrate how very different the information available can be for different groups in the fossil record, I'd like to introduce you to four imaginary friends, all of whom want to analyze paleontological data in a phylogenetic framework (Table 22.1). Brock wants to test if there are trends in body size in planktonic microfossils from the Late Cenozoic, which have an extremely well-sampled fossil record owing to their abundance in deep-sea ocean cores. Misty wants to use comparative methods to test if there is a relationship between body size and taxon longevity in Mid-Paleozoic brachiopods. Lance wants to use phylogenetic data to study the patterns of evolution in biomechanical traits among Mesozoic avians and other theropod dinosaurs. Erika intends to study the rate of evolution of floral traits and already has a well-supported molecular phylogeny for the living members of an angiosperm clade, but she also wants to include information from a few extremely well-preserved specimens from a unique fossil bed in the Mid Cenozoic.

Although some readers might assume I am referring to specific individuals in paleontology with Erika, Lance, Misty, and Brock, I would like to stress that these combinations of groups and research interests were chosen purely out of the rhetorical concern of what best illustrates the issues presented by the fossil record. I do not provide citations here to real analyses similar to these examples because I do not want readers to mistakenly wonder if I am targeting anyone for the deficiencies of their group. I myself study macroevolution in the extinct Graptoloidea, a group of fossil plankton which present their own issues; however, these issues are more specific to graptoloids and thus are less useful for examples in this chapter.

Brock, Misty, Lance, and Erika all have extremely different datasets, because the fossil records they are working with are so very different. All four of them will have to deal with issues not encountered if they were only analyzing a molecular phylogeny of extant taxa. As stressed above, even though they share the same goal of applying comparative methods that require a time-calibrated phylogeny, all four will likely end up using very different solutions to achieve that goal. To understand the inherent differences, we will consider the major issues of the fossil record and discuss how each of our four friends is affected.

Table 22.1 Summary of our hypothetical paleobiology-inclined friends, with their questions and groups of study

Name of hypothetical worker	Research topic	Group of research interest
Brock	Trends in body size evolution	Cenozoic planktonic microfossils
Misty	Relationship between body size and taxon longevity	Mid-Paleozoic brachiopods
Lance	Testing patterns of biomechanical trait evolution	Mesozoic avians and other theropod dinosaurs
Erika	Rate of evolution of floral traits	An extant angiosperm clade with fossil taxa from a locality of exceptional preservation

As discussed throughout this chapter, their group of interest will present various challenges to obtaining a timescaled phylogeny. Thus, their choice of group may create more difficulties and require more research effort than the application of comparative methods to address these different research topics

22.2.1 An Incompletely Sampled Geologic Record

First and foremost, anyone who wishes to analyze anything about the fossil record has to understand the fundamental nature of the fossil record: Everything is incompletely sampled. While sampling issues are certainly important to every field of science, the incomplete gaps and tattered remains of the fossil record are immediately apparent to most observers. To paraphrase a colleague, this has made paleontology the science of studying a degraded biological record. How we deal with this degradation is the art of paleontology.

The key to understanding the incompleteness of the fossil record is stratigraphy. Stratigraphy, for those without a geological background, is the order and position of sedimentary rocks in geologic outcrops (Fig. 22.1). Sedimentary rocks form when sediments such as sand, mud, and silt are deposited in some environment where they become buried, compacted, and eventually lithify (become rock). Fossils are found within units of sedimentary rock, a series of stratigraphic layers that share similar geologic characteristics, probably reflecting a similar environment of deposition (series of related units are called formations). Many units are barren, reflecting environments in which skeletonized organisms did not live or could not be preserved in (or both), with a very small number capable of preserving non-skeletonized organisms. The vast majority of sedimentary units are also highly incomplete, each containing and often separated by innumerous sedimentation hiatuses and intervals of erosion, reflecting the discontinuity of sedimentation rate (Sadler 1981). The formation of sedimentary rocks and their preservation to the present ultimately is a reflection of the ever shifting balance between deposition and erosion, resulting in many regions of the world that completely lack any rock record for a particular interval of time. Finally, there are

many aspects of the relationship between the sediments buried and the fossils we find which reflect the interaction between biotic communities and their environment (see Patzkowsky and Holland 2012, for a comprehensive discussion), and these relationships impact when and how often we sample particular taxa (e.g., Holland 2003).

Thus, the fossil record is incomplete because the rock record is incomplete: Some intervals of time have almost no fossil record available, while for other intervals, the rock record only captures a small number of regions and habitats, with varying degrees of completeness. Except in groups with extremely high sampling potential that lived in environments of high and constant deposition, there will always be some fraction of species missing from the fossil record. Even for taxa sampled in the fossil record, we almost certainly sampled them some unknown time after they originated and an equally unknown time before they went extinct, making their first and last appearances biased estimates for those events. In addition, the sparseness of rocks we can absolutely date means that for most fossil records, first and last appearances are always known with some imprecision (Figs. 22.1 and 22.2).

Knowing that the fossil record is incompletely sampled, the question becomes how we can quantify and account for the uncertainty introduced by that incompleteness (e.g., Strauss and Sadler 1989). In order to understand sampling beyond just assuming it exists, we must choose a model of sampling processes. A commonly assumed and mathematically tractable model is that sampling (both preservation and collection of specimens) is a Poisson process occurring at some time-homogenous rate (e.g., Huelsenbeck and Rannala 1997; Solow and Smith 1997; Foote 1997; Stadler 2010), much like the birth and death models used to study diversification (e.g., Kendall 1948; Raup et al. 1973; Raup 1985; Nee et al. 1992; Foote 2000). A Poisson process is a statistical model of some process that produces a rare event occurring over some interval at a constant rate: The number of events that occur in an interval is described by the Poisson distribution, and the waiting times between events is described by the exponential distribution. Although such a simple, uniform model is almost certainly incorrect, only a few studies have explored the limitations of assuming uniform sampling or have presented alternative models (e.g., Holland 2003; Liow et al. 2010; Wagner and Marcot 2013). Others have split sampling into two separate processes, modeling preservation and collection of preserved specimens individually (Heath 2012). Alternatively, some approaches do not model sampling rate or probability explicitly; instead, sampling intensity is modeled indirectly using various proxies, such as the amount of exposed rock outcrop (Raup 1976; Peters and Foote 2001; Smith and McGowan 2007; Lloyd 2012).

The incomplete sampling of the fossil record is entangled with another basic limit of the rock record: the limits of geologic timescales. The vast majority of fossil finds cannot be assigned with precision to a particular date, but rather to some discrete interval, often several million years long. Some exceptional sedimentary records exist in which events can be pinned with very precise dates as a result of great diligence and the luck to find a very complete rock record, such as
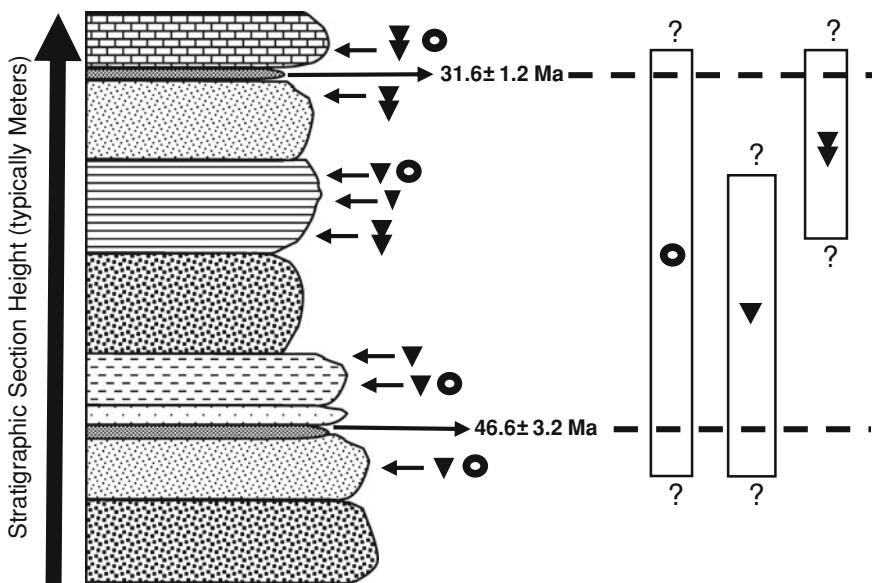
Fig. 22.1 The relationship between the fossil record and the rock record is most obvious at the level of a single hypothetical outcrop, seen here as an idealized stratigraphic section. At this particular locality, horizontal sedimentary rock units of different composition are found stacked vertically, with oldest sediments at the bottom (a common feature of outcrops but not common enough). The *doughnut*, *triangle*, and *double triangle* represent fossils found at different beds, split into separate morphotaxa by differences in their morphology. The collection of taxa in the section is patchy and unevenly distributed, with some units apparently lacking fossils, and some collections within a taxon's range lacking that particular fossil. Atypically, this section has not just one but two *ash beds* that can be absolutely dated, revealing a span of more than 15 million years. As shown from the range bars shown on the *right*, the *double-triangle* taxon is first sampled and the *triangle* taxon is last sampled within this time span, but whether they truly originated/went extinct during that interval is complicated by the incomplete sampling. Omitted from this simplified example is the common association between rock types and faunas and the tendency for rock types to reoccur in a cyclical manner (see Patzkowsky and Holland 2012, for numerous empirical examples)

with many deep ocean cores. Most fossil finds lack the potential to be resolved so finely. This temporal uncertainty represents the lack of biostratigraphic, lithological, and geochemical evidence to further resolve the age of a sedimentary unit. For example, if some fossil outcrop is composed of only long-lived taxa from a rock formation bounded by geochemically obtained absolute dates that are extremely distant, then the age of that outcrop will be very poorly constrained (e.g., Fig. 22.1).

Our ability to resolve the appearance dates of a fossil can differ widely because the stratigraphic quality of each particular locality can be very different, such that resolution can vary greatly within a group of fossil taxa, even among close relatives. One fossil taxon might be known to first occur within a two-million-year-long interval in the Early Triassic, while its sister taxon may be very hard to
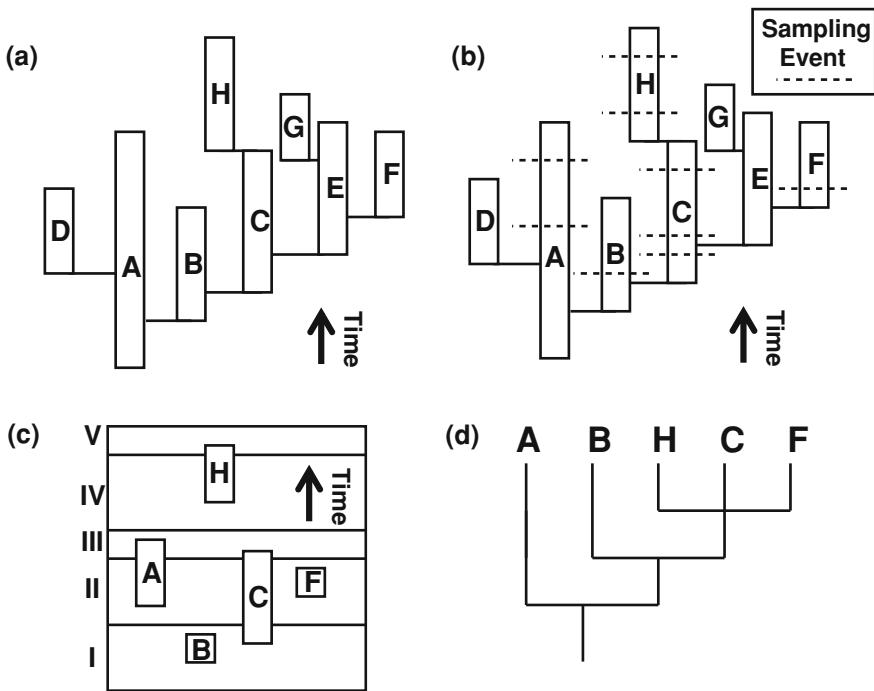
Fig. 22.2 An example model of phylogeny and sampling in the fossil record. **a** In this diagram, a clade of persistent morphotaxa are separated by anagenesis and budding cladogenesis events (Foote 1996). **b** Sampling events are distributed randomly across the evolutionary history of this group. **c** This means only a small proportion of taxa are sampled, and their observed ranges are truncated with respect to their true ranges. Furthermore, we may only be able to resolve first and last appearance dates to within stratigraphic intervals, often of uneven duration. **d** The nested cladogram that would be ideally obtained for the sampled taxa. Patterns of branching and ancestry among taxa can lead to intrinsic polytomies, such as the one containing taxa *C*, *E*, and *H*, which have no correct resolution on an unscaled cladogram (Bapst 2013b)

pin down, maybe known from a single site constrained only as far as being anywhere within the fifty-million-year-long Triassic interval. Furthermore, a given taxon may be sampled many times over within a single interval (as exemplified in Fig. 22.1), to a degree which may be unexpected under homogenous models of sampling. Particularly in the marine fossil record, the number of finds or collections can be so dense that they are often not measured (or possibly not even quantifiable), meaning that the only recorded information is the first and last intervals in which those morphotaxa appear.

Our understanding of the timing of appearances in the fossil record can also differ considerably across groups and environments (e.g., Wagner and Marcot 2013). Occurrences listed from discrete time intervals and the low probability of sampling the true first or last appearance of a given taxon can combine to provide considerable uncertainty and bias in 'when' observed morphotaxa originate and go extinct.

This provides a sometimes unexpected hurdle to those who want to make use of timing information from the fossil record. Marine records tend to be much more complete and known with much better precision than terrestrial records, because of higher and more constant sedimentation rates which allow for more complete rock records. Brock's microfossils will probably have excellent stratigraphic records available from a variety of deep-sea cores, with exceptional dating precision and extremely dense sampling through time. The first and last appearances of Misty's Paleozoic brachiopods will probably be resolved to geologic stages, if not to shorter intervals. Lance's theropods are probably much more incompletely known. Morphospecies in many terrestrial tetrapod groups are known from only a single collection or formation, unlike the long-ranging microfossils and marine invertebrates, suggesting these fossil records are more incomplete and making it difficult to infer the relative stratigraphic order of these collections. Ultimately, the majority of Lance's taxa may be known only from a set of earliest and latest dates for each geologic stage they were collected in. Erika's plant fossils will likely depend on the geologic context of those beds of exceptional preservation they are found in. If Erika is lucky and the fossil bed is perhaps associated with several volcanic ash layers, she may have very specific dates for when these fossils were buried. Otherwise, she might know only that her fossils are from some point in a lengthy interval, perhaps spanning several million years (e.g., the Early Miocene).

22.2.2 The Availability of Paleontological Phylogenies

To apply phylogeny-based approaches, you have to have some sort of branching topology. By necessity, paleontological phylogenies utilize datasets of morphological features to reconstruct relationships. Many issues have been raised with phylogenies inferred from morphological characters, and those seeking to apply phylogenetic comparative methods to fossil data should understand these potential drawbacks (e.g., Rieppel and Kierney 2002; Scotland et al. 2003; Wagner 2000, 2012). Independently of issues with morphological systematics, making a new phylogeny from a new character matrix of morphological data is a time-consuming task that can take a career to do properly. As it is, scientists who want to use phylogeny-based approaches often plan on finding trees in the literature rather than making them entirely anew. Unfortunately, published phylogenetic hypotheses in paleontology often predate modern cladistic practices or are not available for a given set of taxa. Current phylogenetic effort is also unevenly distributed among groups. The relationships of many marine invertebrate groups receive relatively little attention compared to vertebrate groups. For example, Lance's theropods may have an abundance of published cladograms relative to Misty's Paleozoic brachiopods. Among invertebrate groups, somewhat more cladistic effort is been devoted to arthropods, echinoderms, and graptolites than others, based on a per-publication metric (Neige et al. 2009). In some groups, traditional taxonomy alone may be the closest approximation to information on the relationships among taxa.

The methods used to infer relationships also vary considerably among groups, with many complete phylogenies for a fossil group constructed by hand, based on a mélange of stratigraphic order, morphology, and expert opinion, rather than the product of computational algorithms (e.g., Pearson 1998). Additionally, some studies that use an explicit parsimony criterion may not have been applied via optimization algorithms, instead involving the comparison of a small set of candidate topologies by hand (e.g., Fortey and Cooper 1986). Non-computational, hand-drawn trees make up the majority of published topologies for some groups, particularly microfossils. These are sometimes referred to as lineage or 'stratophenetic' phylogenies (Gingerich 1979), although that term is also used for some computational clustering approaches (e.g., Wei 1994; Roopnarine 2005; Hannisdal 2006, 2009). Some microfossil workers (e.g., Aze et al. 2011; Ezard et al. 2011) have preferred stratophenetic phylogenies over results from computational cladistics analyses of molecular and morphological data. This preference is generally argued on the basis that cladistics analyses greatly disagree with other sources of evidence, such as stratigraphic order, or that formal analyses are too under-sampled taxonomically to provide consistent results.

Unfortunately, handmade lineage phylogenies are often much more inclusive of known taxa than trees constructed from formal character matrices and computational analyses and often represent the most taxonomically complete phylogenies. The incompleteness of computational phylogenies is partly an issue of taxonomic level, with such analyses often performed in the fossil record only at the supraspecific level, with genera or families. In other groups, the taxonomic sampling of cladograms is simply patchy because the fossil record leaves only incomplete specimens, and some taxa are much more incomplete than other. Thus, in some fossil groups, a few species may be known in exceptional detail, while the majority of taxa are known from much more incomplete material. The taxa included in phylogenies tend to be the more complete, well-preserved taxa. This is problematic, as preservation potential can vary non-randomly with ecology, morphology, and habitat (e.g., Valentine et al. 2006), such that available cladograms may unwittingly represent biased taxon sampling with respect to these factors. For our intrepid comparative paleobiologists, this can produce biases in any attempt to study the evolution of ecologies and morphologies that correlate with preservation.

Of course, incomplete taxon sampling and a lack of prior phylogenetic work are also an issue in molecular phylogenetics, but the reasons for those biases in the fossil record are very different and so these biases must be evaluated on their own terms. Furthermore, the broad scale of quality and repeatability of what counts as paleontological phylogenetics for many groups is very different from conflicts in molecular biology, such as the issue of whether a mitochondrial tree is comparable to a nuclear gene tree. Ultimately, any worker interested in applying comparative methods in the fossil record will have to decide on their minimum criteria for an acceptable, useable hypothesis of relationships.

To illustrate how different the phylogenetic data available for different groups in the fossil record can be, let us consider what our four friends would find in a stereotypical situation. Lance is probably doing the best: The most detailed species-level phylogenies are often available for tetrapod vertebrates. Misty might locate cladograms made from a computational algorithm, but they might include only a portion of the brachiopods in her dataset: Maybe a species-level cladogram for a single genus or a cladistic analysis conducted at the family or genus level. Brock will likely find only some detailed (but probably handmade) stratophenetic lineage phylogenies for his microfossil group. Finally, Erika may discover that her plant fossil taxa have only been loosely placed in the taxonomic hierarchy and their position never indicated in even the most informal of phylogenies.

22.2.3 A Timescaled Tree

One of the biggest hurdles to any worker looking to use paleobiological data in a phylogeny-based analysis is that the vast majority of such analyses require trees which are scaled to time (or some measure of expected evolutionary change; Garland et al. 1992). Modern biological studies often use dates from molecular clock analyses, calibrated using constraints from the fossil record when available. While many groups in the fossil record have extant members, allowing clade divergence dates to be taken directly from molecular clock analyses, not all of them do. Less obviously, even in those groups that do have extant members, a number of sub-clades will probably lack extant taxa. Without extant members and (ignoring the rare cases with ancient DNA), molecular clock dates do not directly aid the inference of branching times in an extinct clade (although they can indirectly inform other dating approaches for fossils; e.g., Ronquist et al. 2012). Even assuming a phylogeny where every extinct taxon had an extant sister, the node splitting each of these sister pairs cannot be dated under a molecular clock approach. Thus, dates derived from molecular clocks may be difficult to relate to divergence times in a paleontological dataset, unless there is a clear distinction between stem and crown taxa for all living clades.

Existing phylogenies that contain fossil taxa are, with very few exceptions as of the moment, not timescaled. Although methods exist to simultaneously infer relationships among fossil taxa and timescale the resulting phylogeny (e.g., Marcot and Fox 2008; Pyron 2011; Ronquist et al. 2012), they are not yet widely used. Instead, the majority of phylogenies found in the paleontological literature are unscaled cladograms, usually obtained via parsimony analyses. Thus, any phylogeny of fossil taxa constructed or found by searching the literature is likely an unscaled cladogram (like Fig. 22.1d). The divergences among fossil lineages cannot be dated with molecular clocks, and, until recently, there was relatively little information on how to proceed with timescaling the available cladogram with stratigraphic occurrence data. Although the fossil record can be read at face value, this is probably unrealistic in all but the most exquisitely sampled of groups

and brings unwelcome artifacts when more derived taxa appear earlier than more primitive taxa. To deal with these issues, numerous methods were developed to timescale a cladogram and are discussed in detail later in this chapter.

In a stereotypical scenario, three of our four scientists probably will not readily obtain a timescaled tree. If Brock's microfossil taxa were on a stratophenetic tree, then that tree is already timescaled. Misty and Lance probably have cladograms (of some sort), but these cladograms probably would not be timescaled. Instead, they will have to utilize the available stratigraphic information. Assuming Erika has found a phylogeny that includes her plant fossils, she will probably also need to utilize stratigraphic data to obtain divergence dates for extinct taxa or she will be unable to place her fossil taxa on a timescaled phylogeny.

22.2.4 Morphotaxa, Ancestors, and 'Intrinsic' Polytomies

Fossil taxa are almost entirely delimited based on their morphology. While all paleontologists would prefer to work with taxonomic units comparable to biological species, this is not possible for every group. Even ignoring the mercurial nature of species-level taxonomy in some groups, taxa equivalent to biological species sometimes are not distinguishable given the traits that readily preserve. Many groups are analyzed at the lowest taxonomic level that can be consistently recognized by experts: in some groups, that is species, in others, genera. However, even in those groups which can be split to 'species,' it can be difficult to support the claim that such morphospecies represent reproductively isolated 'biological' species, especially given the frequency of cryptic speciation observed in modern studies of some groups (Pfenniger and Schwenk 2007; Trontelj and Fiser 2009) and the morphological variability exhibited by single, reproductively isolated species (Wayne 1986). For groups that are entirely extinct, there may not even be an appropriate modern benchmark to set our expectation of what morphological variation a species might have, meaning that the use of the term 'species' is entirely disconnected from any special connotation that term has when used for living species, for which reproduction isolation can be tested.

Furthermore, the morphotaxa of the fossil record are often not limited to single moments in time. Many morphospecies of marine invertebrate groups, microfossil groups, and even Cenozoic mammals persist across stretches of geologic time (Foote and Raup 1996). Although the fossil specimens may not be perfectly identical, these morphospecies are clearly recognizable from one fossil collection to another. In some cases, discrete morphological characters do not noticeably change over tens of millions of years, as is apparent from the longevity of fossil morphospecies in many invertebrate groups (Eldredge 1971; Van Valen 1973; Stanley 1979; Eldredge et al. 2005). This is at odds with the modern interpretation of a phylogeny in comparative biology, for which taxonomic identity is only appended to terminal tips. These terminal tips represent populations observed at a specific instance in time (usually the present day), not morphologically persistent

lineages that may have been present over several million years. For Misty and Brock, who likely will be dealing with morphospecies that persist across geological time, this issue will need to be dealt with.

At present, we do not have comparative methods that can take into account our observation that morphotaxa can persist without much morphological change. Instead, a particular times of observation has to be selected within each morphotaxon's duration (Bapst 2013a, 2014). The choice of these dates is not arbitrary: They may have a considerable impact on the resulting branch lengths (compare parts d and e in Fig. 22.3, which contrast the use of first and last appearances as the times of observation). Preferably, these times of observation should be the dates corresponding to the specimens used for trait measurement, although this distinction is unhelpful for a static discrete trait. Hunt (2013) took multiple samples from the same morphotaxa, treated as independent observations, to test for relative stasis, but the majority of comparative paleontologists do not collect multiple samples from the same morphospecies across time. Even then, multiple observations of the same discrete trait value across an anagenetic lineage may be unexpected under typical models of discrete trait change and may produce unexpected artifacts of model fitting.

Morphotaxa can also potentially be ancestors to one another. For example, a lineage might be sampled as one morphotaxon and then sampled later as a different morphotaxon, through either anagenetic or cladogenetic change. There may even be multiple intervening morphotaxa that are not sampled, yet the first taxon is still an ancestor of the later ('indirect' ancestry; Foote 1996; for example, taxa C and F in Fig. 22.2). Furthermore, evidence from the fossil record suggests morphotaxa often persist unchanged through speciation events, with analyses supporting a model in which only one daughter lineage accrues immediate morphological change (Wagner and Erwin 1995). This means a single morphotaxon might be the ancestor to a large number of morphologically distinguishable but independently derived descendants. In addition to this, you could have scenarios in which you have sampled an ancestral morphotaxon. Although the incomplete sampling of the fossil record might lead us to predict that the chance of observing ancestors is quite rare, modeling suggests that the possibility of sampling ancestors is quite high (Foote 1996).

Although paleobiologists are not required to make any explicit assumptions about ancestor–descendant relationships in comparative methods, these relationships are likely present and cannot safely be ignored. Many studies ignore the potential for ancestral taxa, often treating the concept of ancestors as philosophically inconsistent with a cladistic motive. In studies which do consider ancestors, one commonly applied rule of thumb treats early appearing, autapomorphy-lacking sister taxa as ancestors (Smith 1994). However, this protocol has been challenged on the grounds that such an assumption ignores the potential for character reversal or loss (Wagner 1996; Polly 1997). Many stratophenetic lineage phylogenies include extensive sets of ancestor–descendant relationships based solely on expert opinion (e.g., Pearson 1998), while a few phylogenetic methods exist that allow ancestors to be inferred based on a compromise of stratigraphy and
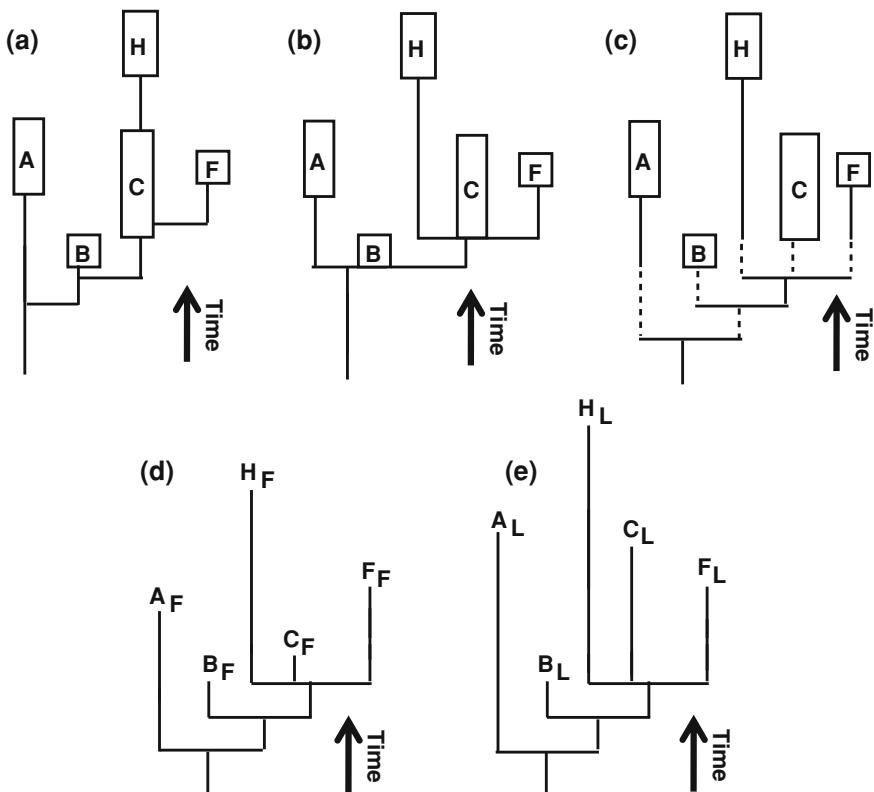
Fig. 22.3 Attempts to timescale phylogenies of fossil taxa for comparative analyses are challenged by the production of ZBLs and the choice of times of observation. **a** The true timescaled phylogeny of the sampled taxon ranges from Fig. 22.2. **b** The timescaled tree inferred given the cladogram and ranges from Fig. 22.2. Taxon B appears to be part of the polytomy containing *C*, *E*, and *H* because of an internal zero-length branch. **c** The basic timescaled tree with all edges constrained to be some minimum branch length, with the resulting extensions shown as *dotted lines*. This both reveals the presence unapparent ZBLs and emulates the minimum branch length timescaling method (Laurin 2004). **d–e** Using the minimum branch length timescaled tree from (**c**), we can contrast the impact on the resulting branch lengths of using the first appearance dates or the last appearance dates as times of observation. Depending on the analyses used, this difference in branch lengths could have an impact on the results of a phylogenetic comparative study (Bapst 2014). In addition, note that these branch lengths differ from those which would be inferred if the true timescaled tree in (**a**) was known

morphology (e.g., Fisher 2008). Ultimately, none of these methods account for the general lack of positive evidence for any particular set of ancestor–descendant relationships. Actually, identifying single ancestors with certainty may be impossible in practical terms. In a phylogenetic comparative framework, ancestral populations can be described in a typical phylogeny structure by placing ancestors on a zero-length branch, effectively attributing zero divergence between a taxon and the lineage that produces later appearing taxa (Bapst 2013a, 2014).

Ancestors have implications beyond their mere existence. If either (a) ancestors are sampled or (b) single morphotaxa have multiple sampled descendants, the expected maximum resolution decreases for a given cladogram of these taxa. This decrease occurs because either scenario introduces true sets of relationships that cannot be reduced to binary branching nodes in a cladogram, producing an 'intrinsic' polytomy (Wagner and Erwin 1995; Bapst 2013b). A lack of intrinsic resolution is particularly expected in well-sampled groups in the fossil record, and thus, we would predict relatively unresolved topologies to result from analyses of such groups. While this introduces another reason to expect phylogenetic uncertainty in paleontological data, the greatest implication is that any polytomy found among a consensus tree could indicate complex scenarios of ancestor–descendant relationships (such as the apparent polytomy in Fig. 22.2d). This possibility should not be ignored when we encounter polytomies in the course of a comparative analysis.

For our four friends, Brock alone probably would have explicit ancestor–descendant relationships from the lineage phylogenies he obtained from the published literature. Misty and Lance probably do not have information on implied ancestor–descendant relationships, although there may be some lineage phylogenies for Misty's brachiopods with such. Erika, who would be lucky to have a phylogeny, will be very unlikely to have any information on ancestor–descendant relationships, although given that her fossil come from a single stratigraphic horizon, perhaps the probability of sampling an ancestor of her extant taxa is negligible.

22.3 Preparing Datasets from the Fossil Record

As with any scientific analysis, paleontological or otherwise, every worker will have to consider whether the available data is sufficient and how to account for uncertainties in any downstream analysis. If the phylogenetic and stratigraphic information available is not sufficient for a scientist to be comfortable with applying a phylogenetic comparative analysis, the only alternative is to do primary data collection and analysis to get better data, like a better phylogeny. Some groups may just be too incompletely known at present to draw deeper understanding about trait evolution in those groups, at least from within a phylogenetic context.

If we decide our data are sufficient, we still have to account for the potential uncertainties listed above: topological relationships, chronostratigraphy, times of observation for persistent taxa, branching times of divergence, and ancestral taxa. Ignoring these similar uncertainties by taking the most likely result to each, or even taking the best solution across all the uncertainties, could be misleading and mask potential uncertainties. For example, under a simple sampling model, the most likely time of extinction for any single taxon is its time of last appearance in the fossil record even though we simultaneously expect, under the same model, that a large proportion of taxa go extinct after they are last observed (Strauss and Sadler 1989).

We need to take such uncertainties into account, not ignore them. For each uncertainty, we can produce multiple likely solutions, and combine them, ultimately expressing our uncertainties as a large number of possible timescaled phylogenies (iterations of this idea can be found in Pol and Norell 2006; Boyd et al. 2011; Bell and Braddy 2012; Lloyd et al. 2012; Sallan and Friedman 2012; Bapst 2013a). We can then analyze across a sample of phylogenies to understand the impact of these various unknowns on an analysis. If our analyses require a timescaled tree, using a sample of multiple phylogenies may be the only avenue to account for the many forms of uncertainty unique to the fossil record.

22.3.1 Combining Taxonomy and Phylogeny

For Misty and her brachiopods, we postulated that perhaps she was not able to find complete phylogenetic information. Perhaps the hypotheses she wishes to test will require collecting character data and inferring new phylogenies, composed of the taxa she is interested in. Yet, there may be a few alternative solutions to consider first. If her main issue is that her taxa are spread across a number of overlapping but conflicting cladistic studies, formal supertree methods have proven to be useful (e.g., Sanderson et al. 1998; Bininda-Emonds et al. 2007; Ruta et al. 2007; Lloyd et al. 2008; Bronzati et al. 2012; Brocklehurst et al. 2013; also see Chap. 3 by Bininda-Emonds in this volume). However, particularly in invertebrate groups where cladistic analyses can be sparse, a different set of issues emerges. Maybe Misty could find trees that cover only a portion of the group she wanted to work on, or the cladograms she found were at the genus level although she aims to test hypotheses at the species level. Phylogenies that include a large majority of taxa may be necessary to avoid results unbiased by taphonomic variation and, yet, such trees are difficult to obtain.

One solution might be to consider the 'gray' phylogenetic data: The relationships implied by stratophenetic diagrams or ancestor–descendant relationships, generally provided in older literature. Similarly, traditional Linnean taxonomy itself is generally assumed to represent a rough first approximation of relationships. The quality of such gray phylogenetic data is extremely variable and needs to be considered carefully.

Scientists who want to have more inclusive datasets of relationships without building new trees must seek compromises between the trees produced using explicit quantitative phylogenetics and the gray phylogenetic literature, including taxonomy. Hybrid 'informal' trees that merge phylogenetic and taxonomic information are becoming increasingly common, particularly among scientists studying patterns of macroevolution and macroecology. Recent examples include the widely used Phylomatic for plant relationships (Webb and Donoghue 2005), which uses phylogenetic relationships when available but defaults to taxonomic

relationships when they are not, or the inclusive, global phylogeny of birds which merges taxonomic and phylogenetic data (Jetz et al. 2012). Some analyses have even used phylogenetic datasets constructed entirely from traditional taxonomic hierarchies (e.g., Green et al. 2011).

If Misty can find a reasonable genus-level tree for her group, she might be able to use taxon lists of genera and species as a way of artificially expanding her phylogeny's taxonomic resolution to the species level. Each genus can be replaced with a polytomy of the species listed for it, reflecting the lack of knowledge for relationships within that taxon.

The main drawback of using taxonomic data as a replacement for phylogeny is that the each higher taxon is assumed to be a valid monophyletic clade. This assumption will need to be considered carefully by each scientist. The monophyly of the traditional taxonomic hierarchy can vary considerably across different fossil groups; for example, Bulman (1970) explicitly defined several form-taxa for graptolites that he considered polyphyletic and many others which were clearly paraphyletic. If we can reject polyphyly but not paraphyly for a group, we can include the possibility of such in our taxonomic–phylogenetic summaries by collapsing the nodes uniting potentially paraphyletic taxa. This effectively produces much larger polytomies, acknowledging the uncertainty in the relationships. However, although this approach is becoming more popular, we badly need analyses comparing phylogenetic trees and taxonomic hierarchies for the same groups and testing their relative capability to address common macroevolutionary questions.

Ultimately, such taxonomic–phylogenetic summaries may become dominated by unresolved polytomies, reflecting the phylogenetic uncertainty of this approach. Phylogenetic uncertainty is certainly not limited to datasets that make use of taxonomic information as a replacement for phylogenetic data; for instance, many paleontological cladograms are also poorly resolved. By randomly resolving polytomies in our dataset many times, we can test whether a particular analysis consistently produces the same result across the potential range of topologies. Unfortunately, some common comparative analyses are positively misled by phylogenetic uncertainty, particularly tests of phylogenetic signal (Davies et al. 2011). Furthermore, such effects may vary from dataset to dataset, depending on the distribution of phylogenetic uncertainty. Approaches such as parametric bootstrapping on phylogenetic datasets (e.g., Boettiger et al. 2012) may be useful for identifying whether phylogenetic uncertainty could be biasing analytical conclusions.

22.3.2 Timescaling in the Fossil Record

Once we have a cladogram containing the taxa we want (maybe not fully resolved), we next have to worry about how we will timescale this topology. This issue is confounded with the need to consider potential ancestor–descendant

relationships and the requirement that tip taxa represent populations, if examined taxa persist over considerable geologic time.

All timescaling methods are dependent on the quality of the stratigraphic data available. Data on temporal occurrences for taxa should, at least, record which intervals taxa first and last occur in and the dates for those intervals. If precise dates are available, or dates for individual samples of the fossil taxa, this information can also be useful, particularly if timescaling involves parameterized models of sampling. Such detailed data may be available in online databases, but often the intervals in which taxa are found and dates for those intervals are found in separate references.

Of course, depending on what sort of tree we are using it may already be timescaled, as is the case with Brock's lineage phylogeny of his microfossil group. Such trees are generally produced by hand and not using computational algorithms or a measure of optimality, but there are now a number of exceptions to this statement. Phylogenetic methods which simultaneously infer topology and timescale the divergences between lineages have increasingly gained popularity over the past two decades. The earliest approaches, sometimes referred to as stratocladistics or strato-likelihood, applied a hybrid of maximum parsimony phylogenetics with an algorithm that could choose less parsimonious phylogenies if they exhibited a better fit to the temporal order of taxa in the fossil record (Fisher 1991, 1994, 2008; Wagner 1998; Marcot and Fox 2008). More recently, it was proposed to infer timescaled phylogenies under both probabilistic models for morphological character change (Lewis 2001) and models of taxonomic sampling in the fossil record, within a likelihood or Bayesian framework (Wagner and Marcot 2010, 2013). A parallel effort has also seen the development of simultaneous methods which use morphological clocks in a Bayesian framework to produce 'tip-dated' timescaled phylogenies of fossil and extant taxa (Pyron 2011; Ronquist et al. 2012; as used in Slater 2013; Wood et al. 2013). In the majority of these applications, an uninformative tree prior is used that allows divergence dates to be mostly informed by the combined molecular and morphological clocks and the first appearance dates, instead of a model of sampling in the fossil record (e.g., Ronquist et al. 2012). Probabilistic models of sampling through time (Stadler 2010) have been implemented for use in estimating viral phylogenies, where samples were taken throughout an epidemic, but were recently applied to datasets containing fossil taxa (Alexandrou et al. 2013). As these Bayesian methods develop further, hopefully adding the ability to assign taxa as ancestral lineages, they will likely become the methods of choice for generating timescaled phylogenies of fossil taxa.

For now, most phylogenetic datasets in paleontology will likely consist of unscaled cladograms, such as the datasets of Misty and Lance. This means that they will need to apply some type of post-inference timescaling algorithm to their cladogram, using their collected stratigraphic data. Most of these methods are fairly ad hoc, with no underlying statistical model. The most basic timescaling method proposes that clades are as old as the oldest taxon within that clade (Norell 1992; Smith 1994; example in Fig. 22.3b). Ancestral taxa are sometimes

inferred, but only if such taxa lack autapomorphies and are sampled earlier than their sister lineage (e.g., Wickstrom and Donoghue 2005).

The greatest issue with this 'basic' method (as I have termed it; Bapst 2013a and 2014) is the presence of artificial zero-length branches ('ZBLs') in the resulting timescaled phylogenies (Hunt and Carrano 2010). These are introduced in two different ways: First, if taxa are known from single occurrences or if the tree is being timescaled so the tip-dates used are the first appearance times, then every sister pair of taxa will produce at least one terminal ZBL, as there is zero time between this earliest first appearance and the branching time for each clade (e.g., the branch leading to taxon *B* in Fig. 22.3c). Secondly, ZBLs are introduced when more derived taxa appear earlier than more primitive taxa in a clade, causing internal branches to be zero-length (e.g., the branch ending in the clade containing taxa *B*, *C*, *E*, and *H* in Fig. 22.3c). Such mismatches between phylogeny and stratigraphy are expected under any scenario of incomplete sampling, but ZBLs pose a number of issues. First, they may pose algorithmic difficulties for some comparative software, but secondly, such ZBLs are generally interpreted by comparative analyses as implying multiple, simultaneous branching events. Taxa placed on a zero-length terminal branch are effectively assumed to represent direct ancestors of their sister lineage (Hunt and Carrano 2010). Several arbitrary transformations to deal with this issue have been used, such as moving branching times earlier in time, such that all branches have some minimum branch length (Laurin 2004; example in Fig. 22.3c).

One timescaling approach assumes a morphological clock and uses the number of character state transitions (estimated during phylogeny inference) as a proxy for the relative length of timescaled branches (e.g., Ruta et al. 2006; Brusatte et al. 2008; Lloyd et al. 2012). A related approach extends branch lengths to maximize the fit to a model of Brownian motion for some continuous trait, such as body size (Laurin 2011). Caution is necessary in applying such strict morphological clock approaches, as considerable evidence exists from the fossil record for the heterogeneity of rates of trait evolution (e.g., Wagner 1995; Bapst et al. 2012; Lloyd et al. 2012). The simultaneous Bayesian methods described above also make use of a morphological clock, but allow for the relaxed clock models implemented previously for molecular data, thus allowing for some heterogeneity of rates of character change (Ronquist et al. 2012).

Some paleontological studies have opted to avoid the need for a timescaled phylogeny for analyses of trait evolution (e.g., Friedman 2009; Pyenson and Sponberg 2011; Smith 2012; Pittman et al. 2013). While some of these studies use this merely for convenience or because stratigraphic information is unavailable or incomplete, other studies argue that if all branches are set to 'unit-length' (i.e., 1, or some other nonzero value), than the resulting tree has branch lengths that simply reflects the evolutionary potential under a 'speciational' model of evolution, as might be expected under punctuated equilibrium. This assumption is somewhat analogous to the explanation given for the commonly used kappa transformation (Pagel 1999), which scales the set of branch lengths in a comparative analysis from their true length to 1, and is thus considered to reflect a measure of

speciational trait change. Unfortunately, the branching nodes of any cladogram composed of extant, extinct, or any combination thereof are probably a poor reflection of the speciation events within a group (Pennell et al. 2014). Extant phylogenies are missing the branching events that lead to extinct lineages, and even phylogenies that include extinct taxa will be missing those unknown lineages that we have never sampled, due to the capricious nature of the fossil record. Given that the necessary biological and geological assumptions are not well supported, using equal branch lengths should be avoided except in analyses where the choice of branch lengths demonstrably has little or no effect on the result.

Very recently, parameterized models of sampling in the fossil record have been used to provide a context in timescaling cladograms (Bapst 2013a; Wagner and Marcot 2013). A considerable advantage of these model-based post-inference methods is that they can consider taxa as potential ancestors based on the calculated probability of sampling ancestors. Unfortunately, these same methods also require a priori estimates, as either rates or probabilities, of sampling (Wagner and Marcot 2013), or extinction and branching in addition (Bapst 2013a). Such estimates are often available for marine invertebrate records, such as Misty's brachiopods, but not for more poorly sampled fossil records, like Lance's theropods. The proposed simultaneous Bayesian methods avoid this issue by simultaneously sampling these parameters simultaneously from the data during analysis (Wagner and Marcot 2010).

A major issue of timescaling in the fossil record is that unless a morphological clock is assumed, there is not any information that allows us to narrow down when a branching event occurred. Similarly, nothing indicates whether a specific taxon is likely to be an ancestor not. Even with divergence dating informed by a morphological clock, uncertainty will be present even if morphology does change at a constant rate, as there will always be only a finite number of morphological characters to measure change in. Thus, considerable uncertainty in divergence times exists in general, even when probabilistic models of sampling and morphology are utilized. As mentioned previously, the majority of such sampling models treat the first and last appearance times of specific taxa as the most likely times that they also originated and went extinct (Strauss and Sadler 1989). For the purposes of inferring branching times, a single set of divergence times obtained via maximum likelihood with such models will be highly misleading. Additionally, it is difficult to utilize the information from confidence intervals on branching times in typical comparative methods, particularly given that divergence dates for nested nodes on a single phylogeny constrain each other, such that more nested (derived) nodes have to occur later in time relative to ancestral nodes.

Bayesian approaches that simultaneously infer topology and timescale allow for the direct consideration of uncertainty by returning a large posterior sample of timescaled phylogenies. A stochastic alternative for post-inference timescaling methods is to generate many timescaled trees, each one with node ages sampled from some probability distribution, with node ages for a single tree sampled sequentially for consistency. Analyses can then be run over the sample of timescaled phylogenies. Tomiya (2013) constrained a small number of divergence

dates using estimates from molecular clock studies and then sampled (and resampled) ages for the remaining nodes under a uniform probability distribution. I recently introduced the cal3 method (Bapst 2013a), which stochastically samples node ages from the probability density implied by a probabilistic model of sampling in the fossil record.

This stochastic approach brings other benefits, beyond just divergence data. The cal3 method extends this stochastic algorithm and the sampling model to stochastically assign ancestral taxa and resolve soft polytomies relative to the amount of stratigraphic misfit expected given the sampling rate (Bapst 2013a). Additionally, the uncertainty in fossil appearance dates from discrete intervals can be accounted for by randomly resampling dates from within those intervals for each generated timescaled phylogeny (e.g., Pol and Norell 2006; Lloyd et al. 2012; Sallan and Friedman 2012). Stochastic timescaling can also account for issues related to the times of observation of morphotaxa. While typical choices have been to use either the first or last appearance times to date terminal tips (Fig. 22.3d–e), these may not be the most reasonable choices in analyses of trait evolution. Ideally, times of observations for such analyses should represent the precise date of the specimens used for measuring traits, but this information is often not available. One alternative is to randomly resample dates from within a taxon's stratigraphic range, like the solution for obtaining dates from discrete intervals. Stochastic timescaling provides a framework for considering all these diverse sources of uncertainty in our analyses.

Bayesian tip-dating or post-inference stochastic methods are ideal approaches to timescaling if the ultimate goal is to apply phylogenetic comparative methods. In a recent study, I compared the performance of the basic method, the minimum branch length transformation and the stochastic cal3 method in a series of simulations where comparative data were generated and then analyzed on trees timescaled using these different methods (Bapst 2014). The non-stochastic timescaling methods were allowed a partial stochastic element by randomly resolving polytomies and resampling dates from within simulated discrete intervals, such that all the analyses were done on samples of multiple timescaled trees. Although some analyses were not particularly affected by the differences in these timescaling methods, this was not true for all comparative analyses. Estimates of the rate of trait evolution and comparisons between models of trait evolution were particularly biased by artifacts introduced by the basic and minimum branch length timescaling methods. This suggests that the choice of an appropriate timescaling method may be just as important for some phylogeny-based analyses as choosing an acceptable cladogram.

Unfortunately, differences in datasets entail differences in methods. At present, only scientists who work on well-sampled fossil records, such as Misty, can utilize the cal3 method and other methods because of the need for detailed information about the sampling intensity of the fossil record. Workers like Lance will have to apply methods that are more arbitrary, while being aware that the methods they use may be introducing unwanted artifacts, like erroneously inferring very short

branches. These timescaling artifacts may propagate errors and biases when applying comparative methods (Bapst 2014).

In the online practical material associated with this chapter (located at http://www.mpcm-evolution.com), I demonstrate the timescaling approaches available in the freely available software package *paleotree* (Bapst 2012) for the statistical programming language R (R Core Team 2013). These examples use a real dataset of retiolitid graptolite genera, consisting of a consensus tree from Bates et al. (2005) and chronological data from Sadler et al. (2009).

22.3.3 Cheating with the Fossil Record: Not Placing Fossil Data Directly on Trees

Clearly, obtaining a timescaled tree of fossil taxa is a difficult and sometimes time-consuming process, but is it absolutely necessary to ask the evolutionary questions we are interested in? For the moment, let us ignore the approaches discussed so far, which require us to have a timescaled phylogeny of fossil taxa.

Fossil data certainly inform us about past patterns, but it is clear that we may not always have sufficient information to place extinct taxa with certainty within a phylogeny, neither topologically nor with respect to when a taxon diverged from other lineages. If we do not feel certain enough about the information at hand, we cannot feel confident to proceed with analyses of evolutionary history. Our friend Erika's flower fossils may simply be known to have features indicative of her study group and nothing else. However, there are ways of including information from the fossil record without having fossil taxa placed as terminals within a tree.

Slater et al. (2012) introduced a promising approach for studying trait evolution by treating fossil data as informative priors in a Bayesian framework. In their method, trait values known from the fossil record only needed to be loosely associated with some clade on a molecular phylogeny of extant taxa. These trait values were then used an informative priors on the trait value at a particular node, essentially using a fossil as an estimate of ancestral morphology. The procedure for setting these priors at nodes is analogous to the application of fossil calibrations as node age priors in molecular clock analyses. Using this approach in both simulations and mammalian fossil data, Slater et al. found that even weakly constrained trait data like this could allow for considerably more power in distinguishing patterns of trait evolution. However, simulations by Slater et al. found that the fossil-prior approach had less power than including a fossil as an actual tip, suggesting that it is still preferable to include fossil taxa as tip taxa if possible.

Approaches, such as this informative prior framework, can avoid placing fossil taxa on a timescaled tree and shall probably become valuable for groups which are diverse in the modern but have poor fossil records. I would recommend this approach for scientists like Erika, who only have some fossils loosely known to represent early representatives of a group much better known from the modern.

However, just like setting fossil calibrations, there has to be reasonable evidence that a fossil specimen represents an early member of a particular extant clade. Hence, the Slater et al.' method does not allow us to consider information from fossils that represent distantly related and/or extinct clades.

22.4 Conclusion

Using the fossil record in comparative analyses is a pursuit that requires considerable care, both with respect to issues common to most of the fossil record, but also for the aspects specific to each group of fossil taxa. The issues of the fossil record cannot be solved without pluralism, as no single workflow can satisfy every scientist seeking to utilize paleontological data. The four individuals we followed had the same motivation but different types of data available, requiring them to adopt very different procedures just to reach the stage where they could apply phylogenetic comparative methods. In the long term, methods will hopefully be developed that can deal with more of the differences among different paleontological datasets, particularly Bayesian phylogenetic inference with a tip-dating approach. Eventually, development of methods might provide a single framework for paleontological comparative analyses, but until that point, forcing a single methodology across datasets will restrict our ability to utilize the advantages of the fossil record.

References

Alexandrou MA, Swartz BA, Matzke NJ, Oakley TH (2013) Genome duplication and multiple evolutionary origins of complex migratory behavior in Salmonidae. Mol Phylogenet Evol 69(3):514–523. doi:http://dx.doi.org/10.1016/j.ympev.2013.07.026

Alfaro ME, Santini F, Brock C, Alamillo H, Dornburg A, Rabosky DL, Carnevale G, Harmon LJ (2009) Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. Proc Natl Acad Sci 106(32):13410–13414

Alroy J (1998) Cope's rule and the dynamics of body mass evolution in North American fossil mammals. Science 280(5364):731–734

Alroy J (2000) Understanding the dynamics of trends within evolving lineages. Paleobiology 26(3):319–329

Aze T, Ezard THG, Purvis A, Coxall HK, Stewart DRM, Wade BS, Pearson PN (2011) A phylogeny of Cenozoic macroperforate planktonic foraminifera from fossil data. Biol Rev 86(4):900–927. doi:10.1111/j.1469-185X.2011.00178.x

Bapst DW (2012) paleotree: an R package for paleontological and phylogenetic analyses of evolution. Methods Ecol Evol 3(5):803–807. doi:10.1111/j.2041-210X.2012.00223.x

Bapst DW (2013a) A stochastic rate-calibrated method for time-scaling phylogenies of fossil taxa. Methods Ecol Evol 4(8):724–733. doi:10.1111/2041-210x.12081

Bapst DW (2013b) When can clades be potentially resolved with morphology? PLoS ONE 8(4):e62312. doi:10.1371/journal.pone.0062312

Bapst DW (2014) Assessing the effect of time-scaling methods on phylogeny-based analyses in the fossil record. Paleobiology 40(3):331–351

Bapst DW, Bullock PC, Melchin MJ, Sheets HD, Mitchell CE (2012) Graptoloid diversity and disparity became decoupled during the Ordovician mass extinction. Proc Natl Acad Sci 109(9):3428–3433

Bates DEB, Kozlowska A, Lenz AC (2005) Silurian retiolitid graptolites: morphology and evolution. Acta Palaeontol Pol 50(4):705–720

Bell MA, Braddy SJ (2012) Cope's rule in the Ordovician trilobite family Asaphidae (order Asaphida): patterns across multiple most parsimonious trees. Hist Biol 24(3):223–230. doi:10.1080/08912963.2011.616201

Benson RBJ, Choiniere JN (2013) Rates of dinosaur limb evolution provide evidence for exceptional radiation in Mesozoic birds. Proceedings of the Royal Society B: Biological Sciences 280(1768)

Benson RBJ, Evans M, Druckenmiller PS (2012) High diversity, low disparity and small body size in Plesiosaurs (Reptilia, Sauropterygia) from the Triassic-Jurassic boundary. PLoS ONE 7(3):e31838. doi:10.1371/journal.pone.0031838

Benton MJ, Donoghue PCJ (2007) Paleontological evidence to date the tree of life. Mol Biol Evol 24(1):26–53

Benton MJ, Hitchin R (1997) Congruence between phylogenetic and stratigraphic data on the history of life. Proc R Soc Lond B: Biol Sci 264(1383):885–890

Benton MJ, Storrs GW (1994) Testing the quality of the fossil record: Paleontological knowledge is improving. Geology 22(2):111–114

Bininda-Emonds ORP, Cardillo M, Jones KE, MacPhee RDE, Beck RMD, Grenyer R, Price SA, Vos RA, Gittleman JL, Purvis A (2007) The delayed rise of present-day mammals. Nature 446(7135):507–512

Boettiger C, Coop G, Ralph P (2012) Is your phylogeny informative? Measuring the power of comparative methods. Evolution 66(7):2240–2251. doi:10.1111/j.1558-5646.2011.01574.x

Boyd CA, Cleland TP, Marrero NL, Clarke JA (2011) Exploring the effects of phylogenetic uncertainty and consensus trees on stratigraphic consistency scores: a new program and a standardized method. Cladistics 27(1):52–60. doi:10.1111/j.1096-0031.2010.00320.x

Brocklehurst N, Kammerer CF, Fröbisch J (2013) The early evolution of synapsids, and the influence of sampling on their fossil record. Paleobiology 39:470–490. doi:10.1666/12049

Bronzati M, Montefeltro FC, Langer MC (2012) A species-level supertree of Crocodyliformes. Hist Biol 24(6):598–606. doi:10.1080/08912963.2012.662680

Brusatte SL, Benton MJ, Ruta M, Lloyd GT (2008) Superiority, competition, and opportunism in the evolutionary radiation of dinosaurs. Science 321(5895):1485–1488

Bulman OMB (1970) Treatise in invertebrate paleontology, Pt. V: Graptolithina, vol Part V. Treatise on invertebrate paleontology. University of Kansas Press and the Geological Society of America, Lawrence, KS

Chan KMA, Moore BR (2002) Whole-tree methods for detecting differential diversification rates. Syst Biol 51(6):855–865

Core Team R (2013) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Davies TJ, Kraft NJB, Salamin N, Wolkovich EM (2011) Incompletely resolved phylogenetic trees inflate estimates of phylogenetic conservatism. Ecology 93(2):242–247. doi:10.1890/11-1360.1

Didier G, Royer-Carenzi M, Laurin M (2012) The reconstructed evolutionary process with the fossil record. J Theor Biol 315:26–37. doi:10.1016/j.jtbi.2012.08.046

Eldredge N (1971) The allopatric model and phylogeny in Paleozoic invertebrates. Evolution 25(1):156–167

Eldredge N, Thompson JN, Brakefield PM, Gavrilets S, Jablonski D, Jackson JBC, Lenski RE, Lieberman BS, McPeek MA, Miller W (2005) The dynamics of evolutionary stasis. Paleobiology 31(sp5):133–145. doi:10.1666/0094-8373(2005)031[0133:TDOES]2.0.CO;2

Evans AR, Jones D, Boyer AG, Brown JH, Costa DP, Ernest SKM, Fitzgerald EMG, Fortelius M, Gittleman JL, Hamilton MJ, Harding LE, Lintulaakso K, Lyons SK, Okie JG, Saarinen JJ, Sibly RM, Smith FA, Stephens PR, Theodor JM, Uhen MD (2012) The maximum rate of mammal evolution. Proc Natl Acad Sci 109(11):4187–4190

Ezard THG, Aze T, Pearson PN, Purvis A (2011) Interplay between changing climate and species' ecology drives macroevolutionary dynamics. Science 332(6027):349–351

Felsenstein J (1988) Phylogenies and Quantitative Characters. Annu Rev Ecol Syst 19(1):445

Finarelli JA, Flynn JJ (2006) Ancestral state reconstruction of body size in the Canifornia (Carnivora, Mammalia): the effects of incorporating data from the fossil record. Syst Biol 55(2):301–313

Fisher DC (1991) Phylogenetic analysis and its implication in evolutionary paleobiology. In: Gilinsky NL, Signor PW (eds) Analytical paleobiology. Paleontological Society, Knoxville, Tennessee, pp 103–122

Fisher DC (1994) Stratocladistics: morphological and temporal patterns and their relation to phylogenetic process. In: Grande L, Rieppel O (eds) Interpreting the hierarchy of nature. Academic Press, San Diego, pp 133–171

Fisher DC (2008) Stratocladistics: Integrating Temporal Data and Character Data in Phylogenetic Inference. Annu Rev Ecol Evol Syst 39(1):365–385

Foote M (1996) On the probability of ancestors in the fossil record. Paleobiology 22(2):141–151

Foote M (1997) Estimating taxonomic durations and preservation probability. Paleobiology 23(3):278–300

Foote M (2000) Origination and extinction components of taxonomic diversity: general problems. In: Erwin DH, Wing SL (eds) Deep time: paleobiology's perspective. The Paleontological Society, Lawrence, Kansas, pp 74–102

Foote M, Raup DM (1996) Fossil preservation and the stratigraphic ranges of taxa. Paleobiology 22(2):121–140

Fortey RA, Cooper RA (1986) A phylogenetic classification of the graptoloids. Palaeontology 29(4):631–654

Friedman M (2009) Ecomorphological selectivity among marine teleost fishes during the end-Cretaceous extinction. Proc Natl Acad Sci 106(13):5218–5223

Fusco G, Garland JT, Hunt G, Hughes NC (2012) Developmental trait evolution in trilobites. Evolution 66(2):314–329. doi:10.1111/j.1558-5646.2011.01447.x

Garland T Jr, Harvey PH, Ives AR (1992) Procedures for the analysis of comparative data using phylogenetically independent contrasts. Syst Biol 41(1):18–32

Gates TA, Prieto-Márquez A, Zanno LE (2012) Mountain building triggered late cretaceous North American megaherbivore dinosaur radiation. PLoS ONE 7(8):e42135. doi:10.1371/journal.pone.0042135

Gingerich PD (1979) The stratophenetic approach to phylogeny reconstruction in vertebrate paleontology. Phylogenet Anal Paleontol 1:41–77

Green WA, Hunt G, Wing SL, DiMichele WA (2011) Does extinction wield an axe or pruning shears? How interactions between phylogeny and ecology affect patterns of extinction. Paleobiology 37(1):72–91. doi:10.1666/09078.1

Guinot G, Adnet S, Cappetta H (2012) An analytical approach for estimating fossil record and diversification events in sharks, skates and rays. PLoS ONE 7(9):e44632. doi:10.1371/journal.pone.0044632

Hannisdal B (2006) Phenotypic evolution in the fossil record: numerical experiments. J Geol 114(2):133–153. doi:10.1086/499569

Hannisdal B (2009) Inferring phenotypic evolution in the fossil record by Bayesian inversion. Paleobiology 33(1):98–115. doi:10.1666/06038.1

Heath TA (2012) A hierarchical Bayesian model for calibrating estimates of species divergence times. Syst Biol 61(5):793–809

Holland SM (2003) Confidence limits on fossil ranges that account for facies changes. Paleobiology 29(4):468–479

Hopkins MJ (2011) How species longevity, intraspecific morphological variation, and geographic range size are related: a comparison using late Cambrian trilobites. Evolution 65(11):3253–3273. doi:10.1111/j.1558-5646.2011.01379.x

Hopkins MJ (2013) Decoupling of taxonomic diversity and morphological disparity during decline of the Cambrian trilobite family Pterocephaliidae. J Evol Biol 26(8):1665–1676. doi:10.1111/jeb.12164

Huelsenbeck JP (1994) Comparing the stratigraphic record to estimates of phylogeny. Paleobiology 20(4):470–483

Huelsenbeck JP, Rannala B (1997) Maximum likelihood estimation of phylogeny using stratigraphic data. Paleobiology 23(2):174–180

Hunt G (2013) Testing the link between phenotypic evolution and speciation: an integrated palaeontological and phylogenetic analysis. Methods Ecol Evol 4(8):714–723. doi:10.1111/2041-210x.12085

Hunt G, Carrano MT (2010) Models and methods for analyzing phenotypic evolution in lineages and clades. In: Alroy J, Hunt G (eds) Short course on quantitative methods in paleobiology, vol 16., Paleontological SocietyNew Haven, Conneticut, pp 245–269

Jetz W, Thomas GH, Joy JB, Hartmann K, Mooers AO (2012) The global diversity of birds in space and time. Nature 491(7424):444–448. http://www.nature.com/nature/journal/v491/n7424/abs/nature11631.html#supplementary-information

Kendall DG (1948) On the generalized "birth-and-death" process. Ann Math Stat 19(1):1–15

Lane A, Janis CM, Sepkoski JJ (2005) Estimating paleodiversity: a test of the taxic and phylogenetic methods. Paleobiology 31(1):21–34

Laurin M (2004) The evolution of body size, Cope's rule and the origin of amniotes. Syst Biol 53(4):594–622

Laurin M (2011) Use of paleontological and phylogenetic data in comparative and paleobiological analyses: a few recent developments. In: Pontarotti P (ed) Evolutionary biology: concepts, biodiversity, macroevolution and genome evolution. Springer, Berlin, pp 121–138. doi:10.1007/978-3-642-20763-1_8

Lewis PO (2001) A likelihood approach to estimating phylogeny from discrete morphological character data. Syst Biol 50(6):913–925

Liow LH, Quental TB, Marshall CR (2010) When can decreasing diversification rates be detected with molecular phylogenies and the fossil record? Syst Biol 59(6):646–659

Lloyd GT (2012) A refined modelling approach to assess the influence of sampling on palaeobiodiversity curves: new support for declining Cretaceous dinosaur richness. Biol Lett 8(1):123–126

Lloyd GT, Davis KE, Pisani D, Tarver JE, Ruta M, Sakamoto M, Hone DWE, Jennings R, Benton MJ (2008) Dinosaurs and the Cretaceous terrestrial revolution. Proc Roy Soc B: Biol Sci 275(1650):2483–2490

Lloyd GT, Wang SC, Brusatte SL (2012) Identifying heterogeneity in rates of morphological evolution: discrete character change in the evolution of Lungfish (Sarcopterygii, Dipnoi). Evolution 66(2):330–348. doi:10.1111/j.1558-5646.2011.01460.x

Losos Jonathan B (2010) Adaptive radiation, ecological opportunity, and evolutionary determinism. Am Nat 175(6):623–639. doi:10.1086/652433

Marcot JD, Fox DL (2008) StrataPhy: a new computer program for stratocladistics analysis. Palaeo-Electronica 11(1):5a

Mooers AØ, Heard SB (1997) Inferring evolutionary processes from phylogenetic tree shape. Q Rev Biol 72(1):31–54

Nee S, Mooers AO, Harvey PH (1992) Tempo and mode of evolution revealed from molecular phylogenies. Proc Natl Acad Sci USA 89(17):8322–8326

Neige P, Brayard A, Gerber S, Rouget I (2009) Les Ammonoïdes (Mollusca, Cephalopoda): avancées et contributions récentes à la paléobiologie évolutive. CR Palevol 8(2–3):167–178

Norell MA (1992) Taxic origin and temporal diversity: the effect of phylogeny. In: Novacek MJ, Wheeler QD (eds) Extinction and phylogeny. Columbia University Press, New York, pp 89–118

Norell MA (1996) Ghost taxa, ancestors, and assumptions: a comment on Wagner. Paleobiology 22(3):453–455

Norell MA, Novacek MJ (1992) The fossil record and evolution: comparing cladistic and paleontologic evidence for vertebrate history. Science 255(5052):1690–1693

Nowak MD, Smith AB, Simpson C, Zwickl DJ (2013) A simple method for estimating informative node age priors for the fossil calibration of molecular divergence time analyses. PLoS ONE 8(6):e66245. doi:10.1371/journal.pone.0066245

Oakley TH, Cunningham CW (2000) Independent contrasts succeed where ancestor reconstruction fails in a known bacteriophage phylogeny. Evolution 54(2):397–405

Pagel M (1999) Inferring the historical patterns of biological evolution. Nature 401(6756):877–884

Patzkowsky ME, Holland SM (2012) Stratigraphic paleobiology: understanding the distribution of fossil taxa in time and space. University of Chicago Press, Chicago, IL

Pearson PN (1998) Speciation and extinction asymmetries in paleontological phylogenies: evidence for evolutionary progress? Paleobiology 24(3):305–335

Pennell MW, Harmon LJ, Uyeda JC (2014) Is there room for punctuated equilibrium in macroevolution? Trends Ecol Evol 29(1):23–32. http://dx.doi.org/10.1016/j.tree.2013.07.004

Peters SE, Foote M (2001) Biodiversity in the Phanerozoic: a reinterpretation. Paleobiology 27(4):583–601

Pfenninger M, Schwenk K (2007) Cryptic animal species are homogeneously distributed among taxa and biogeographical regions. BMC Evol Biol 7(1):121

Pittman M, Gatesy SM, Upchurch P, Goswami A, Hutchinson JR (2013) Shake a tail feather: the evolution of the theropod tail into a stiff aerodynamic surface. PLoS ONE 8(5):e63115. doi:10.1371/journal.pone.0063115

Pol D, Norell MA (2001) Comments on the Manhattan stratigraphic measure. Cladistics 17(3):285–289. doi:10.1111/j.1096-0031.2001.tb00125.x

Pol D, Norell MA (2006) Uncertainty in the age of fossils and the stratigraphic fit to phylogenies. Syst Biol 55(3):512–521

Polly PD (1997) Ancestry and species definition in paleontology: a stratocladistic analysis of Paleocene-Eocene Viverravidae (Mammalia, Carnivora) from Wyoming, vol 30(1). Contributions from the Museum of Paleontology, University of Michigan, pp 1–53

Pyenson N, Sponberg S (2011) Reconstructing body size in extinct crown Cetacea (Neoceti) using allometry, phylogenetic methods and tests from the fossil record. J Mamm Evol 18(4):269–288. doi:10.1007/s10914-011-9170-1

Pyron RA (2011) Divergence time estimation using fossils as terminal taxa and the origins of Lissamphibia. Syst Biol 60(4):466–481

Raia P, Carotenuto F, Passaro F, Piras P, Fulgione D, Werdelin L, Saarinen J, Fortelius M (2013) Rapid action in the Palaeogene, the relationship between phenotypic and taxonomic diversification in Coenozoic mammals. Proc Roy Soc B: Biol Sci 280(1750)

Raup DM (1976) Species diversity in the Phanerozoic: an interpretation. Paleobiology 2(4):289–297

Raup DM (1985) Mathematical models of cladogenesis. Paleobiology 11(1):42–52

Raup DM, Gould SJ, Schopf TJM, Simberloff DS (1973) Stochastic models of phylogeny and the evolution of diversity. J Geol 81:525–542

Rieppel O, Kearney M (2002) Similarity. Biol J Linnean Soc 75(1):59–82. doi:10.1046/j.1095-8312.2002.00006.x

Ronquist F, Klopfstein S, Vilhelmsen L, Schulmeister S, Murray DL, Rasnitsyn AP (2012) A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. Syst Biol 61(6):973–999

Roopnarine PD (2005) The likelihood of stratophenetic-based hypotheses of genealogical succession. Spec Pap Palaeontol 73:143–157

Roy K, Hunt G, Jablonski D (2009) Phylogenetic conservatism of extinctions in marine bivalves. Science 325(5941):733–737

Ruta M, Cisneros JC, Liebrecht T, Tsuji LA, Muller J (2011) Amniotes through major biological crises: faunal turnover among Parareptiles and the end-Permian mass extinction. Palaeontology 54(5):1117–1137. doi:10.1111/j.1475-4983.2011.01051.x

Ruta M, Pisani D, Lloyd GT, Benton MJ (2007) A supertree of Temnospondyli: cladogenetic patterns in the most species-rich group of early tetrapods. Proc Roy Soc B: Biol Sci 274(1629):3087–3095

Ruta M, Wagner PJ, Coates MI (2006) Evolutionary patterns in early tetrapods. I. Rapid initial diversification followed by decrease in rates of character change. Proc Roy Soc B: Biol Sci 273(1598):2107–2111

Sadler PM (1981) Sediment accumulation rates and the completeness of stratigraphic sections. J Geol 89(5):569–584

Sadler PM, Cooper RA, Melchin M (2009) High-resolution, early Paleozoic (Ordovician-Silurian) time scales. Geol Soc Am Bull 121(5–6):887–906

Sallan LC, Friedman M (2012) Heads or tails: staged diversification in vertebrate evolutionary radiations. Proc Roy Soc B: Biol Sci 279(1735):2025–2032

Sanderson MJ, Purvis A, Henze C (1998) Phylogenetic supertrees: assembling the trees of life. Trends Ecol Evol 13(3):105–109

Scotland RW, Olmstead RG, Bennett JR (2003) Phylogeny reconstruction: the role of morphology. Syst Biol 52(4):539–548

Siddall ME (1996) Stratigraphic consistency and the shape of things. Syst Biol 45(1):111–115

Simpson C, Kiessling W, Mewis H, Baron-Szabo RC, Müller J (2011) Evolutionary diversification of reef corals: a comparison of the molecular and fossil records. Evolution 65(11):3274–3284. doi:10.1111/j.1558-5646.2011.01365.x

Slater GJ (2013) Phylogenetic evidence for a shift in the mode of mammalian body size evolution at the Cretaceous-Palaeogene boundary. Methods Ecol Evol 4(8):734–744. doi:10.1111/2041-210x.12084

Slater GJ, Harmon LJ, Alfaro ME (2012) Integrating fossils with molecular phylogenies improves inference of trait evolution. Evolution 66(12):3931–3944. doi:10.1111/j.1558-5646.2012.01723.x

Smith AB (1994) Systematics and the fossil record: documenting evolutionary patterns. Blackwell Scientific, Oxford

Smith AB, McGowan AJ (2007) The shape of the Phanerozoic marine palaeodiversity curve: how much can be predicted from the sedimentary rock record of Western Europe? Palaeontology 50(4):765–774

Smith ND (2012) Body mass and foraging ecology predict evolutionary patterns of skeletal pneumaticity in the diverse "waterbird" clade. Evolution 66(4):1059–1078. doi:10.1111/j.1558-5646.2011.01494.x

Solow AR, Smith W (1997) On fossil preservation and the stratigraphic ranges of taxa. Paleobiology 23(3):271–277

Stadler T (2010) Sampling-through-time in birth-death trees. J Theor Biol 267(3):396–404

Stanley SM (1979) Macroevolution: patterns and process. W. H Freeman & Co., San Francisco

Strauss DJ, Sadler PM (1989) Classical confidence intervals and Bayesian probability estimates for ends of local taxon ranges. Math Geol 21:411–427

Tarver JE, Donoghue PCJ (2011) The trouble with topology: phylogenies without fossils provide a revisionist perspective of evolutionary history in topological analyses of diversity. Syst Biol 60(5):700–712

Tomiya S (2013) Body size and extinction risk in terrestrial mammals above the species level. Am Nat 182(6):E196–E214. doi:10.1086/673489

Trontelj P, Fiser C (2009) Cryptic species diversity should not be trivialised. Syst Biodivers 7(01):1–3

Valentine JW, Jablonski D, Kidwell S, Roy K (2006) Assessing the fidelity of the fossil record by using marine bivalves. Proc Natl Acad Sci 103(17):6599–6604

Van Valen L (1973) A new evolutionary law. Evol Theor 1:1–30

Wagner PJ (1995) Diversity patterns among early gastropods: contrasting taxonomic and phylogenetic descriptions. Paleobiology 21(4):410–439

Wagner PJ (1996) Ghost taxa, ancestors, assumptions, and expectations: a reply to Norell. Paleobiology 22(3):456–460

Wagner PJ (1998) A likelihood approach for evaluating estimates of phylogenetic relationships among fossil taxa. Paleobiology 24(4):430–449

Wagner PJ (2000) The quality of the fossil record and the accuracy of phylogenetic inferences about sampling and diversity. Syst Biol 49(1):65–86

Wagner PJ (2012) Modelling rate distributions using character compatibility: implications for morphological evolution among fossil invertebrates. Biol Lett 8(1):143–146

Wagner PJ, Erwin DH (1995) Phylogenetic patterns as tests of speciation models. In: Erwin DH, Anstey RL (eds) New approaches to speciation in the fossil record. Columbia University Press, New York, pp 87–122

Wagner PJ, Erwin DH (2006) Patterns of convergence in general shell form among Paleozoic gastropods. Paleobiology 32(2):316–337. doi:10.1666/04092.1

Wagner PJ, Marcot JD (2010) Probabilistic phylogenetic inference in the fossil record: current and future applications. In: Alroy J, Hunt G (eds) Short course on quantitative methods in paleobiology, vol 16., Paleontological SocietyNew Haven, Connecticut, pp 189–211

Wagner PJ, Marcot JD (2013) Modelling distributions of fossil sampling rates over time, space and taxa: assessment and implications for macroevolutionary studies. Methods Ecol Evol 4(8):703–713. doi:10.1111/2041-210x.12088

Warnock RCM, Yang Z, Donoghue PCJ (2012) Exploring uncertainty in the calibration of the molecular clock. Biol Lett 8(1):156–159

Wayne RK (1986) Cranial morphology of domestic and wild canids: the influence of development on morphological change. Evolution 40(2):243–261. doi:10.2307/2408805

Webb CO, Donoghue MJ (2005) Phylomatic: tree assembly for applied phylogenetics. Mol Ecol Notes 5(1):181–183. doi:10.1111/j.1471-8286.2004.00829.x

Wei K-Y (1994) Stratophenetic tracing of phylogeny using SIMCA pattern recognition technique: a case study of the late Neogene Planktic Foraminifera Globoconella clade. Paleobiology 20(1):52–65

Wickström L, Donoghue PCJ (2005) Cladograms, phylogenies and the veracity of the conodont fossil record. Spec Pap Palaeontol 73:185–218

Wills MA (1999) Congruence between phylogeny and stratigraphy: randomization tests and the gap excess ratio. Syst Biol 48(3):559–580

Wills MA, Barrett PM, Heathcote JF (2008) The modified gap excess ratio (GER*) and the stratigraphic congruence of dinosaur phylogenies. Syst Biol 57(6):891–904

Wood HM, Matzke NJ, Gillespie RG, Griswold CE (2013) Treating fossils as terminal taxa in divergence time estimation reveals ancient vicariance patterns in the Palpimanoid spiders. Syst Biol 62(2):264–284

Zanno LE, Makovicky PJ (2013) No evidence for directional evolution of body mass in herbivorous theropod dinosaurs. Proc Roy Soc B: Biol Sci 280(1751)

Index