László Zsolt Garamszegi
Editor

Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology

Concepts and Practice

Springer

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

L. Z. Garamszegi (ed.), *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology*, DOI: 10.1007/978-3-662-43550-2_2,

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.

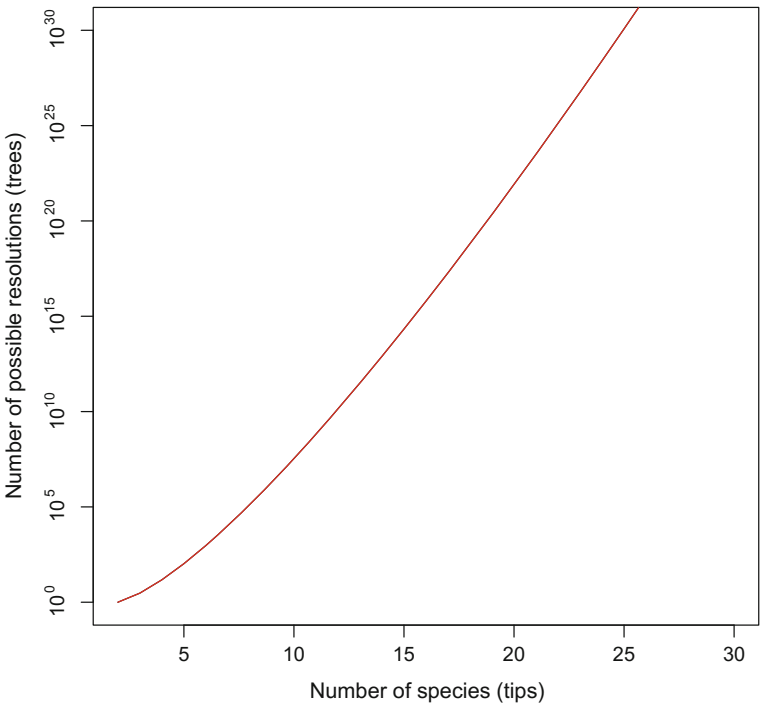

Fig. 2.1 The number of possible phylogenetic resolutions (trees) as a function of the number of species (tips)

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.

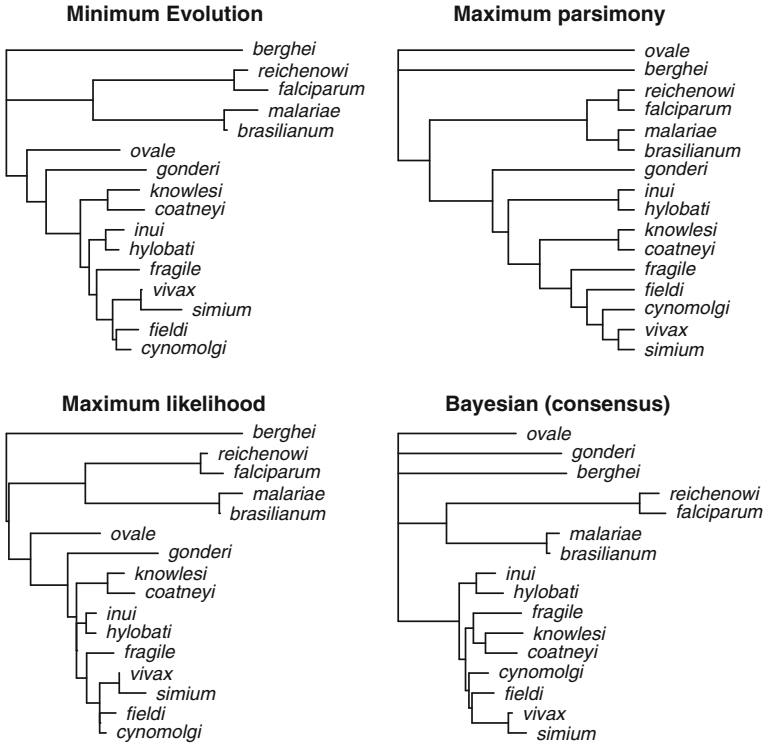

Fig. 2.2 Phylogenetic resolutions of 15 primate-infecting *plasmodium* (malaria) species, as revealed by the analysis of 18S rDNA sequences by different phylogeny estimation methods. The underlying data obtained form the GenBank and correspond to Leclerc et al. (2004)

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.org). 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
	Testing the effect of particular clade on the phylogeny	TreeView TreeEdit 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; Binindа-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