CHAPTER 1

# ESTIMATING SPECIES TREES: AN INTRODUCTION TO CONCEPTS AND MODELS

*L. Lacey Knowles*
*Laura S. Kubatko*



## 1.1 INTRODUCTION

The estimation of relationships among species in an evolutionary context broadly falls within the purview of the discipline of systematics. However, as the central framework in evolutionary (and some ecological) study, the enormous impact of this single endeavor—phylogenetic estimation—is unquestionable. How, and whether, species relationships are accurately inferred are, consequently, issues of broad and far-reaching concern.

The goal of this book is to provide an overview of several recently developed methods for phylogenetic estimation that focus explicitly on the challenges and strengths inherent in the analysis of multilocus data while giving practical guidelines on implementing these approaches. Decreased sequencing costs and increased access to primer sets enhance the relative ease of data collection, providing unprecedented amounts of multilocus sequence for molecular phylogenetic analysis across all of biodiversity (e.g., Goldman and Yang 2008; Hughes et al. 2006; Wiens et al. 2008). Detailed suggestions and discussion throughout the chapters focus on both conceptual and methodological issues, addressing such topics as how results should be interpreted and how to recognize the signs of a problem with an analysis. The combination of theoretical and empirical studies contained herein serves to identify both the strengths and the limitations of these new methods under not only idealized situations with simulated data but also with empirical sequence data. The guidelines also serve to draw attention to the impact that sampling design, marker choice, and taxon sampling will have on the performance of the new methods.

### 1.1.1 Different Tree Types and Their Relationship to Phylogeny

As a characterization of the history of species divergence (including both the pattern and relative timing of lineage splitting), a phylogeny is a tree where both the topology and branch lengths portray information about the evolutionary history of species (Fig. 1.1). While molecular data predominate the pursuit of estimating the evolutionary history of species, the trees estimated from DNA sequences are clearly distinct from, and are not

*Estimating Species Trees: Practical and Theoretical Aspects*, Edited by L. Lacey Knowles and Laura S. Kubatko

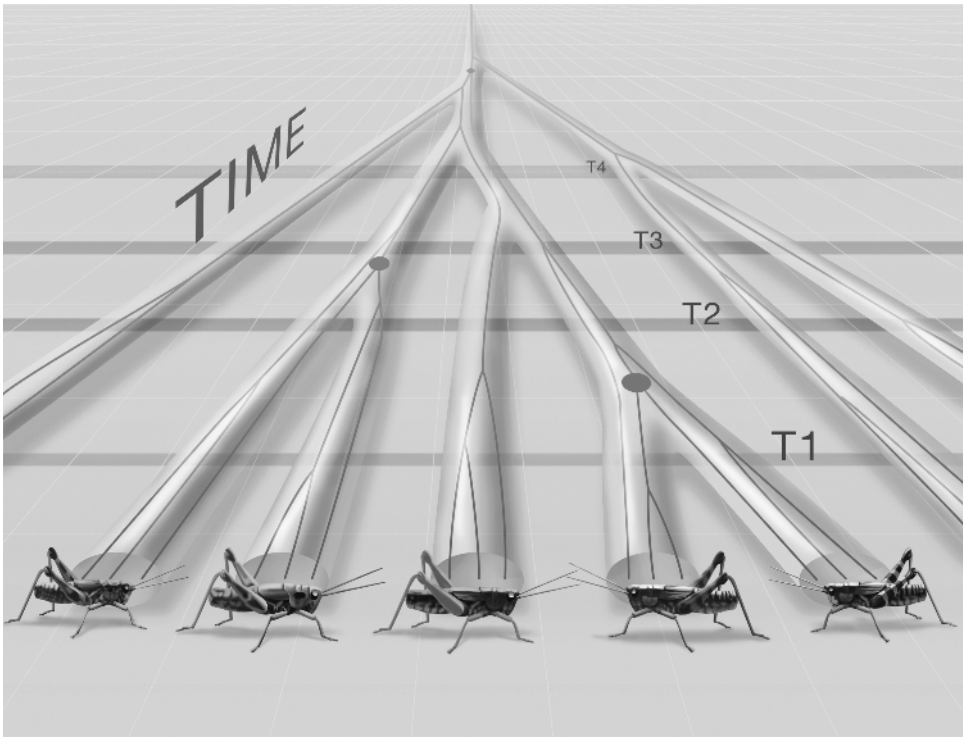


Figure 1.1 Species trees contain information on both the pattern (topology) and timing (branch lengths) of species diversification. This phylogenetic history can be inferred from the gene trees that are embedded within the species lineages, which may or may not be concordant with the species tree (e.g., the deep coalescence of gene lineages marked with the red dots). By incorporating a model of gene lineage coalescence (in addition to the models of nucleotide substitution), the phylogenetic history of species (i.e., the species tree) can be estimated, despite widespread incomplete lineage sorting (i.e., sequences from multiple individuals per species—three individuals for this locus in this case—do not form monophyletic clades). (Illustration by John Megahan.)

synonymous with, the underlying species history—the species tree (Maddison 1997; Slowinski and Page 1999). In contrast to the differing genealogical histories (i.e., gene trees) that might characterize a locus (or a nonrecombining DNA fragment), there is only one species history, whether that history is strictly bifurcating (i.e., a species tree) or involves reticulations, which may or may not obscure species relationships.

The patterns of similarity and differences in the DNA sequences of organisms related by descent from common ancestors implicitly contain information about species relationships. That is, there is an intimate link between gene trees and the species tree in which they are embedded. This link means that gene trees are informative about species phylogenies, yet it is clear that a gene tree should not be equated with a species phylogeny since the evolutionary processes that determine the structure of gene trees differ from those governing species trees. The structure of a species tree is determined by the process of speciation, extinction, and in some cases, hybridization, whereas the gene tree structure reflects not only the proliferation and loss of species lineages but also the population genetic process of mutation and gene lineage coalescence within species lineages, and in some cases, the locus-specific effects of migration between species lineages.

Enormous attention has been dedicated to understanding the theoretical and computational challenges associated with estimating gene trees from molecular data, as well as the practical complications that arise with empirical investigations. For example, in addition to the development of very sophisticated methods for estimating a gene tree from DNA sequences (e.g., accommodating complex models of nucleotide evolution and evaluating the full probability of the data for a set of tree topologies and branch lengths; reviewed in Felsenstein 2004), the impact of various data properties on tree accuracy is also well studied (e.g., the number of base pairs analyzed and taxon sampling; Flynn et al. 2005; Graybeal 1998; Rannala et al. 1998; Rosenberg and Kumar 2001; Wiens 2003; Zwiki and Hillis 2002). In contrast, we are only beginning to understand the theoretical and computational challenges, as well as the practical complications of empirical data, when the target is to obtain an estimate of the species tree. For example, multiple processes may determine the relationship between species and their contained loci (e.g., gene lineage coalescence alone or in combination with gene flow). Moreover, the collection of possible bifurcating trees (i.e., the *tree space*) becomes enormous even for a moderate number of species. For example, even if only bifurcating processes are considered, and ignoring differences in branch lengths, there are approximately $2 \times 10^6$ trees for 10 taxa. The difficulties posed by such issues, as well as strategies for contending with these challenges, are discussed in the following sections that trace the steps from species tree estimation back to the collection of DNA sequence data.

While much of the research on obtaining direct estimates of species trees has been driven by computational developments, these methodological changes do not represent the inception of new core phylogenetic concepts. The recent advances (paradoxically) provide a practical means of returning to the systematic tradition of estimating species relationship. Thus, in spite of the fact that estimating species trees involves a fundamental shift in how molecular data are used and interpreted, the target is still the phylogeny. Estimation of a species tree, in addition to putting the focus on the object of systematic interest, also provides a framework for studying the processes generating a set of contained gene trees because of the explicit distinction between the species tree and gene trees. For example, the discord among gene trees may be biologically meaningful (as opposed to being due to tree-building errors, for example; Jeffroy et al. 2006). The different gene trees may provide insights about the diversification process (e.g., the population size of the taxa relative to the divergence time separating speciation events, or the extent of gene flow among taxa), or whether species trees are meaningful if there is significant horizontal gene transfer, a question that requires empirical evaluation (e.g., Galtier and Daubin 2008).

## 1.2 THE RELATIONSHIP BETWEEN GENE TREES AND SPECIES TREES

Gene trees and species trees are different from one another for a variety of reasons. The most important of these is the possibility that evolutionary processes such as horizontal gene transfer, hybridization, gene duplication, or incomplete lineage sorting lead to differences in the underlying histories of each gene for a given species phylogeny. Understanding these evolutionary processes and their effect on the relationship between gene trees and species trees is thus a problem of central importance to the development of methods for estimating species phylogenies: the goal is estimation of species trees; the data available to do this come in the form of DNA sequences arising from the histories of individual genes. We must therefore strive to understand and effectively model the

process by which sequence data arise on the individual gene trees, conditional on the overall species-level relationships.

The methods described and illustrated in this book incorporate one or more of the evolutionary processes mentioned above, and many of these models are common to several of the subsequent chapters on species tree estimation. For this reason, we will devote the next few sections to giving a relatively broad overview of the common models used to relate gene trees to species trees, with ample references to which the reader is directed to obtain a more detailed explanation. Section 1.2.1 defines the processes of horizontal gene transfer, gene duplication, hybridization, and incomplete lineage sorting, and briefly describes their effects on relationships between gene trees and species trees. Section 1.2.2 gives a more detailed description of the coalescent process because it is fundamental to several of the methods included in this book (e.g., Chapters 2, 4, 5, and 6). Section 1.3 then builds on this by describing methods for modeling nucleotide sequence evolution along gene trees.

### 1.2.1 Evolutionary Mechanisms for Gene Tree Discord

Maddison (1997) provides a very comprehensive description of the processes mentioned below, with explicit discussion of the effects of these processes on individual gene histories. Here we provide the following brief descriptions:

- *Horizontal gene transfer* is a term used to describe a process by which genetic material is transferred from one species to another at a given point in time (thus corresponding to genetic exchange that occurs "horizontally" across a phylogeny), rather than from parent to offspring (which occurs "vertically" on a phylogeny). This could happen, for instance, when a vector such as a virus carries DNA from one species to another and this genetic material is subsequently integrated into the genome of the infected organism. Horizontal gene transfer events are known to occur commonly in the bacteria (Medigue et al. 1991; Syvanen 1994; Valdez and Pinero 1992). Horizontally transferred genes will, at least initially, be more closely related to the ancestors of the organism from which they were derived than to those in which they currently reside, thus leading to gene trees that differ from the species tree.
- *Gene duplication* refers to the event that a copy of a particular gene is inserted into the genome, followed by the subsequent (and separate) evolution of the two copies. If a single copy of the gene is sampled from each organism, the sampling of a duplicated gene might result in the observation of a gene tree that differs from the species tree. Gene duplication events are prevalent in plants, fish, and insects.
- *Hybridization* between species occurs when two distinct species interbreed, with the resulting formation of hybrid organisms that share some genetic material from each of the parental organisms. When hybridization occurs without formation of a new taxonomic lineage that is distinct from the parental lineages from which it was formed, the process is often referred to as *introgression* or *introgressive hybridization*. Hybridization is ubiquitous in nature, with current estimates that approximately 25% of plants and 10% of animals hybridize (Mallet 2007).
- *Incomplete lineage sorting* occurs when multiple gene lineages persist through speciation events. Following a speciation event, some forms of the gene may be lost, while others are maintained and continue to evolve. This process is illustrated in Figure 1.2a, which shows a species tree for three taxa (outlined in bold, black lines) with several embedded gene trees (thinner, colored lines). For example, in the green gene tree, gene lineage C fails to find a most recent common ancestor with gene lineage B during time interval $t$, and instead finds a most recent common ancestor with gene lineage A above the root of the species tree. This leads to a gene tree that

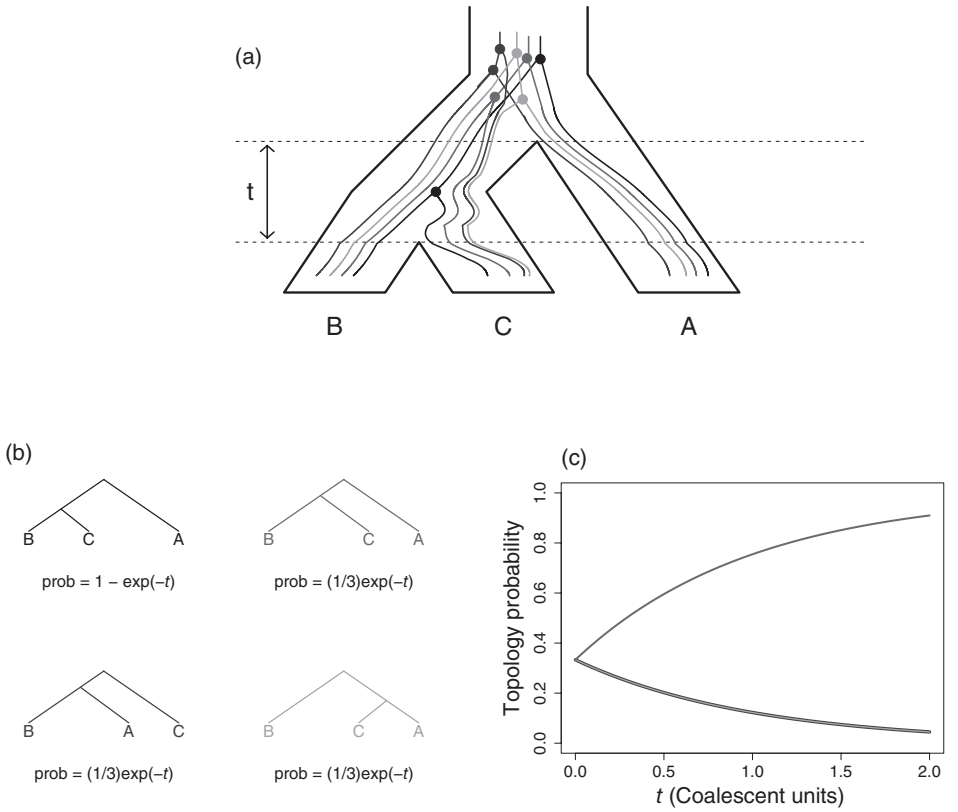


Figure 1.2 Topology probabilities under the coalescent model for three-taxon trees. (a) The species tree is shown outlined in black. The time interval between the two speciation events is $t$, and should be interpreted in coalescent units (number of $2N$ generations). The four embedded trees are the four possible gene histories when deep coalescent events are allowed. (b) The four possible gene histories from (a) are shown separately, with their probabilities under the coalescent model given beneath. Note that the two gene histories in the first row are the same when only the topology is considered, so that the probability of this gene tree topology under the coalescent model is the sum of these two probabilities. Thus, there are only three distinct gene tree topologies in the three-taxon case. (c) Probabilities of each of three gene tree topologies under the coalescent model as a function of the interval of time between speciation events, $t$. Note that the "blue" and "green" gene trees always have the same probabilities. Note also that as $t$ increases, the probability of the "red" gene tree (which is the gene tree with the same topology as the species tree) approaches 1.

differs from the species tree (Fig. 1.2b). It is clear that the possibility of such events can result in gene trees that differ in substantial and important ways from the species tree. This process is commonly modeled by the coalescent.

### 1.2.2 The Coalescent Process and Gene Tree Distributions

Several of the chapters included in this volume develop methodologies for species tree estimation that utilize the coalescent process as a model for the relationship between gene trees and a species tree. For this reason, we include here a more detailed introduction to the basic ideas underlying this process. Excellent books on this topic include the recent works of Hein et al. (2004) and Wakeley (2009).

The *coalescent*, or the *coalescent process*, refers to a mathematical model for the random joining of sampled gene lineages as they are followed back in time. In most

settings in population genetics and phylogenetics, the coalescent process is the model that results from consideration of the large-sample approximation of common population genetic models, such as the Wright–Fisher model and the Moran model (Wakeley 2009). This large-sample approximation is commonly referred to as Kingman's coalescent, in recognition of the derivation of these limiting properties by Kingman in a series of papers (Kingman 1982a, 1982b, 1982c). Other important early works on the coalescent include Tavaré (1984), Takahata and Nei (1985), Pamilo and Nei (1988), Takahata (1989), and Rosenberg (2002).

Under the coalescent model, times to coalescent events follow an exponential distribution. The parameter of the exponential distribution depends on both the number of sampled lineages and the size of the underlying population. For example, consider two lineages sampled from a population of size $2N$. Under the coalescent model, the probability that the two lineages *coalesce* (i.e., find a common ancestor) no more than $t$ units of time into the past is given by $1 - e^{-t/(2N)}$ (this expression follows from the exponential distribution). Probabilities associated with coalescent events involving more than two lineages are given in several places (e.g., Rosenberg 2002).

Using these calculations, one can compute probabilities associated with coalescent events within the branches of a species tree. Because each species tree branch represents an independently evolving population, these branches can be linked together to allow calculation of probability distributions associated with the gene tree topology (Degnan and Salter 2005; Pamilo and Nei 1988; Rosenberg 2002). Note that here (and throughout the book), we use the term *topology* to refer to the shape of the gene tree without regard to the branch lengths. *Gene tree* will refer to the gene tree topology *and* its associated branch lengths. Figure 1.2 provides an example of the calculation of gene tree topology probabilities for a fixed species tree in the case of three taxa. In part (a), the species is shown by the bold, outlined tree, and the four possible gene histories are depicted by thinner, colored lines embedded within the tree. Part (b) shows these four gene histories separately, with their corresponding probabilities under the coalescent model given below. These probabilities follow from the exponentially distributed times to coalescence, as described above. The parameter $t$ refers to the time interval between speciation events, and is given in *coalescent units*, which are the number of $2N$ generations. For example, if $N$ is 100,000, then $t = 0.1$ corresponds to 20,000 generations. Since the gene tree topologies in the first row of part (b) are the same (the two histories differ only in the timing of the coalescent event), the probability of this gene tree topology is the sum of the probabilities of each gene history. Thus, there are a total of three distinct gene trees possible in the three-taxon case. In part (c), the probability of each gene tree topology is plotted as a function of $t$. When $t$ is 0, all three topologies are equally likely since this essentially corresponds to a trifurcating species tree. As $t$ increases, the topology that matches the species tree becomes increasingly probable, with probability very near 1 once the interval between speciation events is larger than approximately 2.0 coalescent units. This occurs because the probability that two gene lineages coalesce in time that is smaller than 2.0 coalescent units is very large under the exponential model. Since only gene lineages B and C are available to coalesce within this interval, the gene trees generated in this way must match the underlying species tree.

The example coalescent model depicted in Figure 1.2 has been extended in several ways. Degnan and Salter (2005) developed an algorithm for computing the probability associated with any gene tree topology given any species tree with speciation times when only a single gene lineage is sampled for each species. In this book, an extension of this methodology to the case in which multiple gene lineages per species are sampled is given in Chapter 4.

The gene tree topology distribution is relevant to the problem of estimating species trees in several ways. First, it allows computation of a likelihood function that can be used to infer the species tree (see, e.g., Carstens and Knowles 2007). Second, the distribution of gene tree topologies for a given species tree under the coalescent has been useful in studying the performance of traditional phylogenetic inference when applied to multilocus data. For example, early reports seemed to indicate that concatenation of multilocus data into a single contiguous stretch of DNA followed by analysis with traditional parsimony, maximum likelihood (ML), or Bayesian methodology designed for a single locus resulted in robust, highly supported phylogenies (e.g., Chen and Li 2001; Gadagkar et al. 2005; Rokas and Carroll 2005; Rokas et al. 2003). However, subsequent studies demonstrated cases in which such procedures could fail (Carstens and Knowles 2007; Kolaczkowski and Thornton 2004; Kubatko and Degnan 2007; Mossel and Vigoda 2005). Consideration of the gene tree topology distribution makes clear one possible reason for the potential poor performance of traditional phylogenetic methods on concatenated sequences, namely, that the concatenation procedure assumes that all data conform to a single gene tree, while in reality each gene has its own history arising within the common species tree.

While the gene tree topology distribution has been vital in understanding and modeling the relationships between gene trees and species trees, it does ignore potential additional information in the gene trees in the form of the branch lengths. Rannala and Yang (2003) provided an important advance in phylogenetic coalescent methodology by explicitly deriving the probability density of gene trees (both topology and branch lengths) given a species tree. They used this distribution to develop a procedure for estimating speciation times and effective population sizes along a fixed species tree for multilocus data. However, the gene tree density that they derived has been used in a variety of other contexts since then. Its usefulness results from the fact that it allows the likelihood function of the species tree given a collection of gene trees to be evaluated. It has therefore been used to develop algorithms that estimate species trees from either a collection of gene trees (Kubatko et al. 2009; Liu et al. 2009; Mossel and Roch 2010) or a set of sequence alignments for a multilocus data set (Edwards et al. 2007; Liu and Pearl 2007). Several of these techniques have been included in this volume (see Chapters 2 and 6).

### 1.2.3 Phylogenetic Extensions of the Coalescent Model

In addition to methods for phylogenetic inference that include the coalescent, some attention has been given to methods that incorporate other evolutionary process in addition to the coalescent. For example, the gene tree distributions described in the previous section make the assumption that there is no gene flow between species following speciation. However, gene flow may be fairly common, particularly between sister species in time intervals just after speciation. Methods that combine the coalescent process with a model that incorporates gene flow have been developed (Hey and Nielsen 2004, 2007; Nielsen and Wakeley 2001) and are in fairly widespread use. However, they are limited at present in that they assume a known species tree and focus instead on estimation of associated population genetic demographic parameters.

Several authors have also considered the problem of integrating the process of hybridization or other horizontal transfer into a coalescent framework. The methods considered range from using simulation to compare expectations under models that do or do not include hybridization or horizontal transfer (e.g., Buckley et al. 2006; Joly et al. 2009; Maureira-Butler et al. 2008) to those that develop explicit models for such processes (e.g., Meng and Kubatko 2009; Nakhleh 2010; Than et al. 2007). This is currently an active area of research, and an example is described in Chapter 6.

## 1.3 THE RELATIONSHIP BETWEEN SEQUENCE DATA AND GENE TREES

Given a model for which gene trees are generated from an underlying species tree (e.g., the coalescent), we now consider the process by which DNA sequence data arise along a gene tree, and the implications that this has for the estimation of gene trees. In fact, this is the situation that is generally more familiar to those working in the field of phylogenetics, and so here only a brief review of the basic ideas is given. For a general review of these ideas, the reader is referred to the book by Felsenstein (2004).

### 1.3.1 Modeling DNA Sequence Evolution along a Gene Tree

All methods based on the likelihood function, which include both the ML and Bayesian methods that will be described below, incorporate some model for the evolution of DNA sequences along a gene tree over time. These models, commonly referred to as nucleotide substitution models, describe the process by which one nucleotide sequence mutates to another over time. The most commonly used models are continuous-time homogeneous Markov processes that satisfy the condition of time reversibility. These models are specified by a matrix that describes the mean instantaneous rate of change from one nucleotide to another. The most general of these models, called the general time-reversible (GTR) model, is one that allows for differences in the rates of all possible nucleotide changes while still satisfying the condition of time-reversibility. Various submodels follow from the GTR model by specifying constraints on the types of changes that happen at varying rates. For example, the simplest model, the Jukes–Cantor model, assumes that all nucleotide substitutions occur at the same rate (see, e.g., Goldman 1993 for a review of commonly used models).

In modeling the process of nucleotide substitution, two other features of molecular evolution are generally incorporated. First, it is well known that different sites in the DNA sequence may evolve at different rates. Thus, it is common to include a model for variation in the rate of evolution across sites. The most common model in this regard is the discrete gamma approximation of Yang (1994). Second, it is common for many sites within a particular gene to be invariant, and thus a parameter that allows a proportion of the sites to be unchanging is commonly included as well.

Given the wide range of possible nucleotide substitution models, a first step in the analysis of most empirical data is an evaluation of which model provides the best fit for a set of aligned DNA sequences. This is generally done through the use of some model selection procedure. Two of the most commonly used procedures are: the use of a series of likelihood ratio tests for nested models, and comparison via the Akaike information criterion (AIC; Akaike 1974) and the Bayesian information criterion (BIC; Schwarz 1978); both are implemented in the package ModelTest (Posada and Crandall 1998). Another popular option is the use of the decision-theoretic criterion as employed in the package DTModSel (Minin et al. 2003). It is generally now fairly widely accepted that each locus should be modeled with its own evolutionary model, so this procedure must be carried out for all loci individually prior to species tree inference.

## 1.4 STATISTICAL INFERENCE OF SPECIES TREES

Thus far, we have introduced the common models that relate DNA sequence data to gene trees, and gene trees to the species tree. Armed with these models, we now return to our goal of developing methods for inferring species trees from multilocus sequence data sets. If we view this inference problem in a statistical framework, the possible methods avail-

able can be broadly classified into two types of approaches: ML methods and Bayesian methods. A feature shared by both of these methods is their utilization of the likelihood function for a species tree as a formal part of the inference procedure. However, the methods differ fundamentally in how they view the inference problem. In the ML framework (generally called the frequentist framework in Statistics), it is assumed that there is a single true underlying species phylogeny. Estimation of this true underlying species phylogeny is then carried out by searching the large space of all possible species phylogenies for the tree or trees that maximize the likelihood function. More detail concerning possible likelihood functions used in species tree inference is given in Section 1.4.1.

In the Bayesian framework, parameters (such as the species tree) are believed to be observations from an underlying distribution, and attention is focused on estimating this distribution. Thus, in the Bayesian setting, the likelihood function is used as part of a probability calculation based on Bayes' theorem to obtain a probability distribution over species trees. In practice, this distribution cannot be computed exactly, and statistical methods to estimate this distribution are required. We give more explicit details of this framework in Section 1.4.2.

In addition, the problem of inferring a species tree can be considered in a parsimony setting, in which one seeks the species tree that is most parsimonious with an observed collection of gene trees. One criterion proposed in this framework is the minimize deep coalescences (MDC) method (Maddison and Knowles 2006), in which the species tree which requires the smallest number of deep coalescent events is preferred. A computationally efficient method for finding species trees under the MDC criterion is described in Chapter 5.

### 1.4.1 ML

As described above, there are two key steps in obtaining ML species tree estimates: first, it must be possible to evaluate the likelihood function for a multilocus data set given a particular species tree; second, a method for searching the space of possible species trees for the particular tree that maximizes this likelihood must be developed. Once the first step is possible, the second problem is identical to that typically considered in phylogenetic estimation of gene trees. Thus, we do not discuss it further here (for additional details, see Felsenstein 2004).

The likelihood function for a multilocus data set can be formulated in several different ways, depending on what type of data is available for inference. For example, a likelihood function for a species tree can be computed from either gene tree topologies, gene trees (both topology and branch lengths), or the aligned nucleotide sequences. In all of these cases, the primary assumption of the methods presented here is that the variability in gene trees arises solely from the coalescent model in the absence of other forces such as gene flow. In addition, it is assumed that loci are sampled so that their gene trees are independent, conditional on the species tree. These assumptions allow formulation of various likelihood functions, which we briefly describe:

- *Likelihood of the species tree given a sample of gene tree topologies*: A possible likelihood for this setting can be written down under the additional assumption that gene tree topologies are known with certainty. This likelihood is computed by multiplying gene tree topology probabilities, which are computed as in Figure 1.2 (see Section 1.2.2; see also Degnan and Salter 2005).
- *Likelihood of the species tree given a sample of gene trees (topology as well as branch lengths)*: In this case, we assume that both gene tree topology *and* branch lengths are known with certainty. Since we deal with gene tree topologies as well as branch lengths, we replace the topology probabilities as computed in Figure 1.2

with the gene tree density as given by Rannala and Yang (2003). This is the likelihood function that is the basis of the Species Trees estimation using Maximum likelihood (STEM) method (Kubatko et al. 2009).

- *Likelihood of the species tree given sequence data for multiple loci*: In additional to those listed above, the assumptions underlying this likelihood function are that sequence data arise according to one of the nucleotide substitution models described above. This likelihood has been written down by Maddison (1997) and Felsenstein (2004), among others. It is computationally intensive to compute this likelihood, as it requires evaluation of a high-dimensional integral over the set of possible gene trees.

One other likelihood function that commonly plays a role in species tree estimation methodology, although it deals with gene trees rather than the species tree, is the following:

- *Likelihood of the gene tree given the DNA sequence data for that gene*: This is the "traditional" likelihood used in single-gene phylogenetics, and is computed under one the models of nucleotide sequence evolution described above. The reader is referred to Felsenstein (2004) for details concerning the computation.

In terms of their use in inferring the species tree, each of these likelihood functions can be used either alone or in combination to produce a criterion for estimation of the species trees. Each such method will have its own strengths and weaknesses, and an important goal of the present volume is to bring to the forefront the particular successes and remaining challenges associated with utilization of the various species tree likelihood functions for phylogenetic inference.

### 1.4.2 Bayesian Analysis

As described above, Bayesian methods are concerned with the study of the *posterior probability distribution* of the parameters of interest, which is obtained using both the *prior distribution* of the parameter and the likelihood of the observed data. The prior distribution represents any known information about the distribution of the parameters of interest before data are collected. The likelihood represents the information about the parameters contained in the data, and is dependent on a particular model that relates the data to the parameters. The prior distribution and the likelihood are used together to compute the posterior distribution according to Bayes' theorem. Intuitively, the idea is that there is some knowledge of likely values of the parameters of interest before data are collected; this is the prior information. Then, data are collected and the likelihood of the data for various hypothesized parameters values can be evaluated. Using this information, the knowledge concerning the parameters of interest is summarized by the posterior distribution, using information from both the prior information and the likelihood of the observed data.

In a phylogenetic setting, the species phylogeny, including the branch lengths, is the parameter of primary interest, although many other parameters might be studied as well (e.g., parameters in the nucleotide substitution models may also be of interest). Bayes' theorem in this settings gives

$$P(Tree, Model|Data) = \frac{P(Data|Tree, Model)\,P(Tree, Model)}{P(Data)}$$

where $P(Data|Tree,Model)$ is the likelihood function and $P(Tree,Model)$ is the prior distribution of the tree and associated model parameters. Because the unconditional probabil-

ity of the data, $P(Data)$, cannot be easily computed (it would involve a summation over all possible tree topologies and an integral over branch lengths in each), a computational technique called Markov chain Monte Carlo (MCMC) is used to approximate the posterior distribution. This estimated posterior distribution can then be used to study whatever features are of most interest. When the species tree estimate is of primary interest, the most common result of a Bayesian phylogenetic analysis is the construction of a consensus tree from the set of trees sampled by the MCMC procedure.

Recent years have seen a dramatic increase in the use of Bayesian methodology for inferring phylogenies. One reason for this is that the Bayesian framework has some important advantages over frequentist methods, including the possibility of carrying out inference for models that include a large number of parameters (which is made feasible by the use of MCMC), as well as the estimation of an entire posterior distribution (rather than a single phylogeny) that can be used to study many aspects of the evolutionary history of the species under study, beyond simple estimation of the species tree. Useful reviews of Bayesian phylogenetic inference have been given by Huelsenbeck and Bollback (2001) and Huelsenbeck et al. (2001) (for a more general review of Bayesian methods in genetics, see Beaumont and Rannala 2004).

## 1.5 COLLECTING DNA SEQUENCE DATA

Obtaining an estimate of a species tree, by definition, requires the collection of multiple loci per species. However, the amount of data needed for obtaining an accurate species tree estimate will differ depending on the history of species divergence. Likewise, how the total sampling effort should be split between sequencing more loci versus individuals will differ (Maddison and Knowles 2006; McCormack et al. 2009). As a consequence, although there may be some general sampling guidelines (e.g., the utility of sampling multiple individuals is limited to recently diverged taxa, assuming divergence has occurred without gene flow), it is not defensible to advocate the collection of vast amounts of DNA sequences, irrespective of the specific details of the species history, at least for the purpose of estimating phylogeny (of course, there may be other legitimate reasons for collecting large-scale genomic data). These issues, as well as an approach for devising species-specific informed sampling strategies, are discussed in Chapter 10.

As the input to methodological approaches for estimating species trees, the quality of gene tree estimates as well as the sequence data from which they are derived impacts the accuracy of species tree estimates. In addition to the computational considerations discussed above, there are a number of other basic data assumptions that cannot be overlooked. These include stochastic tree-building errors (as discussed above) that arise from insufficient sequence length or limited mutational variation, as well as nongenealogical patterns of descent caused by recombination. As a consequence, all sequences used for species tree estimation should first be examined to determine whether they conform to these assumptions. Tests for recombination can be carried out with a number of different of programs, including the four-gamete test, or estimates of per-site recombination rate (e.g., with the program SITES; Hey and Wakeley 1997). For distinguishing between mutational and recombinational variance when some sites might have experienced multiple substitutions, a simulation approach can be used to determine whether estimates of the per-site recombination rate calculated for each locus differs significantly from values obtained from data simulated with no recombination under the estimated model of sequence evolution for each locus (see Knowles and Carstens 2007). For nuclear loci, alleles must also be phased (i.e., to determine which polymorphisms at multiple heterozygous sites occur on the

two allelic copies). Polymerase chain reacation (PCR) subcloning, as well as algorithmic approaches (e.g., analysis with the program PHASE, Stephens and Donnelly 2003), can be used to phase alleles, and are often used in a complementary fashion (e.g., Carstens and Knowles 2007). PCR subcloning can also be used for verifying that the loci are indeed single copy, thereby assuring that only orthologs are used to estimate the species tree.

## 1.6 CONCLUSIONS

Obtaining direct estimates of species trees, as opposed to equating reconstructed gene trees with the phylogenetic history of species, represents an intriguing shift in phylogenetic study. Such a transition has far-reaching implications for how species trees are estimated, expanding the role of explicit models of evolutionary character change to encompass the process of substitution within gene lineages (e.g., Felsenstein 1981; Hennig 1966), as well as the sorting among gene lineages (e.g., Edwards and Cavalli-Sforza 1964; Takahata 1989). With the discovery that discordant gene trees retain a significant signal of phylogenetic history (Maddison and Knowles 2006), what was once just a theoretical possibility (Maddison 1997) is becoming a reality (e.g., Belfiore et al. 2008; Brumfield et al. 2008; Carstens and Knowles 2007; Edwards et al. 2007; Knowles and Carstens 2007; Kubatko et al. 2009; Liu and Pearl 2007; Liu et al. 2008).

The methods discussed in this book represent a fundamental transformation in how gene trees are used and interpreted, whereby what we want to capture—species relationships and divergence times—rather than the type of data we collect, motivates the methodological procedures. As such, the transition toward species tree estimation in phylogenetic study is generally desirable, and broadly applicable. No matter how accurately a gene tree might be estimated, an inescapable biological reality remains—gene trees differ among loci, and they may not match the underlying species tree, making the current methodologies insufficient and ineffectual at handling the DNA sequences now being collected from multiple loci. For example, to make inferences about the processes underlying observed differences in gene trees across loci, a framework in which the expected patterns can be evaluated is essential (e.g., in order to determine whether the different gene trees are consistent with the variance expected given mutational and coalescent processes, or whether some other process like gene glow might be acting). There are also evolutionary histories that defy resolution when gene trees are interpreted literally as the species phylogenetic history, namely, recent species divergence and evolutionary radiations. For these settings, the transition to a species tree perspective is vital in making phylogenetic estimation feasible.

In the following chapters, the combination of empirical investigation, simulation, and theory is used to illustrate the intriguing promises of species tree estimation, while also drawing attention to the limitations and difficulties commonly encountered in practice. With this in mind, the book will serve both as a guide for achieving accurate species tree estimates and as an indicator of areas in need of further work that will hopefully inspire future development.

## REFERENCES

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19:716–723.

Beaumont, M. A. and B. Rannala. 2004. The Bayesian revolution in genetics. *Nature Reviews. Genetics* 5:251–261.

Belfiore, N. M., L. Liu, and C. Moritz. 2008. Multilocus phylogenetics of a rapid radiation in the genus Thomomys (Rodentia: Geomyidae). *Systematic Biology* 57:294–310.

Brumfield, R., L. Liu, D. Lum, and S. V. Edwards. 2008. Comparison of species tree methods for reconstructing the phylogeny of bearded manakins (Aves: Pipridae: *Manacus*) from multilocus sequence data. *Systematic Biology* 57(5):719–731.

Buckley, T., M. Cordeiro, D. Marshall, and C. Simon. 2006. Differentiating between hypotheses of lineage sorting and introgression in New Zealand cicadas (*Maoricicada dugdale*). *Systematic Biology* 55:411–425.

Carstens, B. C. and L. L. Knowles. 2007. Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from *Melanoplus* grasshoppers. *Systematic Biology* 56:400–411.

Chen, F. and W. Li. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *American Journal of Human Genetics* 68:444–456.

Degnan, J. H. and N. A. Rosenberg. 2006. Discordance of species trees with their most likely gene trees. *PLoS Genetics* 3:762–768.

Degnan, J. and L. Salter. 2005. Gene tree distributions under the coalescent process. *Evolution* 59:24–37.

Edwards, A. F. W. and L. L. Cavalli-Sforza. 1964. Reconstruction of evolutionary trees. Phenetic and phylogenetic classification. Eds. V. H. Heywood and J. McNeill. London: Syst. Assoc. Publ. No. 6:67–76.

Edwards, S., L. Liu, and D. Pearl. 2007. High-resolution species trees without concatenation. *Proceedings of the National Academy of Sciences of the United States of America* 104:5936–5941.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum-likelihood approach. *Journal of Molecular Evolution* 17:368–376.

Felsenstein, J. 2004. *Inferring Phylogenies*. Sunderland, MA: Sinauer Associates.

Flynn, J. J., J. A. Finarelli, S. Zehr, J. Hsu, and M. A. Nedbal. 2005. Molecular phylogeny of the Carnivora (Mammalia): assessing the impact of increased sampling on resolving enigmatic relationships. *Systematic Biology* 54:317–337.

Gadagkar, S. R., M. S. Rosenberg, and S. Kumar. 2005. Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. *The Journal of Experimental Zoology* 304B:64–74.

Galtier, N. and V. Daubin. 2008. Dealing with incongruence in phylogenomic analyses. *Philosophical Transactions of the Royal Society of London. Series B. Biological Sciences* 363:4023–4029.

Goldman, N. 1993. Statistical tests of models of DNA substitution. *Journal of Molecular Evolution* 36:182–198.

Goldman, N. and Z. Yang. 2008. Introduction. Statistical and computational challenges in molecular phylogenetics and evolution. *Philosophical Transactions of the Royal Society of London. Series B. Biological Sciences* 363:3889–3892.

Graybeal, A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? *Systematic Biology* 47:9–17.

Hein, J., M. H. Schierup, and C. Wiuf. 2004. *Gene Genealogies, Variation, and Evolution: A Primer in Coalescent Theory*. Oxford: Oxford University Press.

Hennig, W. 1966. *Phylogenetic Systematics*. (English translation). Urbana: University of Illinois Press.

Hey, J. and R. Nielsen. 2004. Mulilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of drosophila pseudoobscura and d. persimilis. *Genetics* 167:747–760.

Hey, J. and R. Nielsen. 2007. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences of the United States of America* 104:2785–2790.

Hey, J. and J. Wakeley. 1997. A coalescent estimator of the population recombination rate. *Genetics* 145: 833–846.

Huelsenbeck, J. P. and J. P. Bollback. 2001. Application of the likelihood function in phylogenetic analysis. In D. J. Balding, M. Bishop, and C. Cannings, eds. *Handbook of Statistical Genetics*. New York: Wiley, pp. 415–444.

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

Hughes, C. E., R. J. Eastwood, and C. D. Bailey. 2006. From famine to feast? Selecting nuclear DNA sequence for plant species-level phylogeny reconstruction. *Philosophical Transaction of the Royal Society Series B* 361:211–225.

Jeffroy, O., H. Brinkmann, F. Delsuc, and H. Philippe. 2006. Phylogenomics: the beginning of incongruence? *Trends in Genetics* 22:225–231.

Joly, S., P. A. McLenachan, and P. J. Lockhart. 2009. A statistical approach for distinguishing hybridization and incomplete lineage sorting. *The American Naturalist* 174:E54–E70.

Kingman, J. F. C. 1982a. Exchangeability and the evolution of large populations. In G. Koch and F. Spizzichino, eds. *Exchangeability in Probability and Statistics*. Amsterdam: North-Holland, pp. 97–112.

Kingman, J. F. C. 1982b. On the genealogy of large populations. *Journal of Applied Probability* 19A:27–43.

Kingman, J. F. C. 1982c. The coalescent. *Stochastic Processes and Their Applications* 13:235–248.

Knowles, L. L. and B. C. Carstens. 2007. Estimating a geographically explicit model of population divergence. *Evolution* 61:477–493.

Kolaczkowski, B. and J. W. Thornton. 2004. Performance of maximum parsimony and maximum likelihood phylogenetics when evolution is heterogeneous. *Nature* 431:980–984.

Kubatko, L. and J. Degnan. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology* 56:17–24.

Kubatko, L., B. Carstens, and L. Knowles. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics* 25(7):971–973.

Liu, L. and D. Pearl. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Systematic Biology* 56:504–514.

Liu, L., D. K. Pearl, R. Brumfield, and S. V. Edwards. 2008. Estimating species trees using multiple allele data. *Evolution* 62:2080–2091.

Liu, L., L. Yu, and D. Pearl. 2009. Maximum tree—a consistent estimator of the species tree. *Journal of Mathematical Biology* 60(1):95–106.

Maddison, W. P. 1997. Gene trees in species trees. *Systematic Biology* 46:523–536.

Maddison, W. P. and L. L. Knowles. 2006. Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology* 55:21–30.

Mallet, J. 2007. Hybrid speciation. *Nature* 446:279–283.

Maureira-Butler, I. J., B. E. Pfeil, A. Muangprom, T. C. Osborn, and J. J. Doyle. 2008. The reticulate history of Medicago (Fabaceae). *Systematic Biology* 57:466–482.

McCormack, J., H. Huang, and L. L. Knowles. 2009. Maximum likelihood estimates of species trees: how accuracy of phylogenetic inference depends upon the divergence history and sampling. *Systematic Biology* 58(5):501–508.

Medigue, C., T. Rouxel, P. Vigier, A. Henaut, and A. Danchin. 1991. Evidence for horizontal gene transfer in *Escherichia coli* speciation. *Journal of Molecular Biology* 222:851–856.

Meng, C. and L. S. Kubatko. 2009. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. *Theoretical Population Biology* 75:35–45.

Minin, V., Z. Abdo, P. Joyce, and J. Sullivan. 2003. Performance-based selection of likelihood models for phylogeny estimation. *Systematic Biology* 52:674–683.

Mossel, E. and S. Roch. 2010. Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 7(1):166–171.

Mossel, E. and E. Vigoda. 2005. Phylogenetic MCMC algorithms are misleading on mixtures of trees. *Science* 309:2207–2209.

Nakhleh, L. 2010. Evolutionary phylogenetic networks: models and issues, In L. Heath and N. Ramakrishnan, eds. *The Problem Solving Handbook for Computational Biology and Bioinformatics*. Springer (forthcoming).

Nielsen, R. and J. Wakeley. 2001. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* 158:885–896.

Pamilo, P. and M. Nei. 1988. Relationships between gene trees and species trees. *Molecular Biology and Evolution* 5:568–583.

Posada, D. and K. A. Crandall. 1998. MODELSTEST: testing the model of DNA substitution. *Bioinformatics* 14:817–818.

Rannala, B. and Z. Yang. 2003. Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. *Genetics* 164:1645–1656.

Rannala, B., J. P. Huelsenbeck, A. Yang, and R. Neilsen. 1998. Taxon sampling and the accuracy of large phylogenies. *Systematic Biology* 47:702–710.

Rokas, A. and S. B. Carroll. 2005. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. *Molecular Biology and Evolution* 22:1337–1344.

Rokas, A., B. Williams, N. King, and S. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425:798–804.

Rosenberg, N. A. 2002. The probability of topological concordance of gene trees and species trees. *Theoretical Population Biology* 61:225–247.

Rosenberg, M. S. and S. Kumar. 2001. Incomplete taxon sampling is not a problem for phylogenetic inference. *Proceedings of the National Academy of Sciences* 98:10751–10756.

Schwarz, G. E. 1978. Estimating the dimension of a model. *Annals of Statistics* 6:461–464.

Slowinski, J. and R. D. Page. 1999. How should species phylogenies be inferred from sequence data? *Systematic Biology* 48:814–825.

Stephens, M., and P. Donnelly. 2003. A comparison of Bayesian methods for haplotype reconstruction. *American Journal of Human Genetics* 73:1162–1169.

Syvanen, M. 1994. Horizontal gene transfer: evidence and possible consequences. *Annual Review of Genetics* 28:237–261.

Takahata, N. 1989. Gene genealogy in three related populations: consistency probability between gene and population trees. *Genetics* 122:957–966.

Takahata, N. and M. Nei. 1985. Gene genealogy and variance of interpopulational nucleotide differences. *Genetics* 110:325–344.

Tavaré, S. 1984. Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoretical Population Biology* 26:119–164.

Than, C., D. Ruths, H. Innan, and L. Nakhleh. 2007. Confounding factors in HGT detection: statistical error, coalescent effects, and multiple solutions. *Journal of Computational Biology* 14:517–535.

Valdez, A. M. and D. Pinero. 1992. Phylogenetic estimation of plasmid exchange in bacteria. *Evolution* 46:641–656.

Wakeley, J. 2009. *Coalescent Theory: An Introduction*. Greenwood Village, CO: Roberts and Co.

Wiens, J. J. 2003. Missing data, incomplete characters and phylogenetic accuracy. *Systematic Biology* 52:528–538.

Wiens, J. J., C. A. Kuczynski, S. A. Smith, D. Mulcahy, J. W. Sites, T. M. Townsend, and T. W. Reeder. 2008. Branch lengths, support, and congruence: testing the phylogenomic approach with 20 nuclear loci in snakes. *Systematic Biology* 57:420–431.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution* 39:306–314.

Zwicki, D. J. and D. M. Hillis. 2002. Increased taxon sampling greatly reduces phylogenetic error. *Systematic Biology* 51:588–598.