Part I

Theory

1 Introduction

This book is part science, part technical, and all about the computational analysis of heritable traits: things like genes, languages, behaviours and morphology. This book is centred around the description of the theory and practice of a particular open source software package called BEAST (Bayesian evolutionary analysis by sampling trees). The BEAST software package started life as a small science project in New Zealand but it has since grown tremendously through the contributions of many scientists from around the world, chief among them the research groups of Alexei Drummond, Andrew Rambaut and Marc Suchard. A full list of contributors to the BEAST software package can be found on the BEAST GitHub page or printed to the screen when running the software.

Very few things challenge the imagination as much as does evolution. Every living thing is the result of the unfolding of this patient process. While the basic concepts of Darwinian evolution and natural selection are second nature to many of us, it is the detail of life's tapestry which still inspires an awe of the natural world. The scientific community has spent a couple of centuries trying to understand the intricacies of the evolutionary process, producing thousands of scientific articles on the subject. Despite this Herculean effort, it is tempting to say that we have only just scratched the surface.

As with many other fields of science, the study of biology has rapidly become dominated by the use of computers in recent years. Computers are the only way that biologists can effectively organise and analyse the vast amounts of genomic data that are now being collected by modern sequencing technologies. Although this revolution of data has really only just begun, it has already resulted in a flourishing industry of computer modelling of molecular evolution.

This book has the modest aim of describing this still new computational science of evolution, at least from the corner we are sitting in. In writing this book we have not aimed for it to be comprehensive and gladly admit that we mostly focus on the models that the BEAST software currently supports. Dealing, as we do, with computer models of evolution, there is a healthy dose of mathematics and statistics. However, we have made a great effort to describe in plain language, as clearly as we can, the essential concepts behind each of the models described in this book. We have also endeavoured to provide interesting examples to illustrate and introduce each of the models. We hope you enjoy it.

1.1 Molecular phylogenetics

The informational molecules central to all biology are deoxyribonucleic acid (DNA), ribonucleic acid (RNA) and protein sequences. These three classes of molecules are commonly referred to in the molecular evolutionary field as *molecular sequences*, and from a mathematical and computational point of view an informational molecule can often be treated simply as a linear sequence of symbols on a defined alphabet (see Figure 1.1). The individual building blocks of DNA and RNA are known as *nucleotides*, while proteins are composed of 20 different *amino acids*. For most life forms it is the DNA double-helix that stores the essential information underpinning the biological function of the organism and it is the (error-prone) replication of this DNA that transmits this information from one generation to the next. Given that replication is a binary reaction that starts with one genome and ends with two similar if not identical genomes, it is unsurprising that the natural and appropriate structure for visualising the replication process over multiple generations is a bifurcating tree. At the broadest scale of consideration the structure of this tree represents the relationships between species and higher-order taxonomic groups. But even when considering a single gene within a single species, the ancestral relationships among genes sampled from that species will be represented by a tree. Such trees have come to be referred to as *phylogenies* and it is becoming clear that the field of *molecular phylogenetics* is relevant to almost every scientific question that deals with the informational molecules of biology. Furthermore, many of the concepts developed to understand molecular evolution have turned out to transfer with little modification to the analysis of other types of heritable information in natural systems, including language and culture. It is unsurprising then that a book on computational evolutionary analysis would start with phylogenetics.

The study of phylogenetics is principally concerned with reconstructing the evolutionary history (phylogenetic tree) of related species, individuals or genes. Although algorithmic approaches to phylogenetics pre-date genetic data, it was the availability of genetic data, first allozymes and protein sequences, and then later DNA sequences, that provided the impetus for development in the area.

A phylogenetic tree is estimated from some data, typically a multiple sequence alignment (see Figure 1.2), representing a set of homologous (derived from a common ancestor) genes or genomic sequences that have been aligned, so that their comparable regions are matched up. The process of aligning a set of homologous sequences is itself a hard computational problem, and is in fact entangled with that of estimating a phylogenetic tree (Lunter et al. 2005; Redelings and Suchard 2005). Nevertheless, following convention we will – for the most part – assume that a multiple sequence alignment is known and predicate phylogenetic reconstruction on it.

DNA	{A,C,G,T}
RNA	{A,C,G,U}
Proteins	{A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y}

Figure 1.1 The alphabets of the three informational molecular classes.

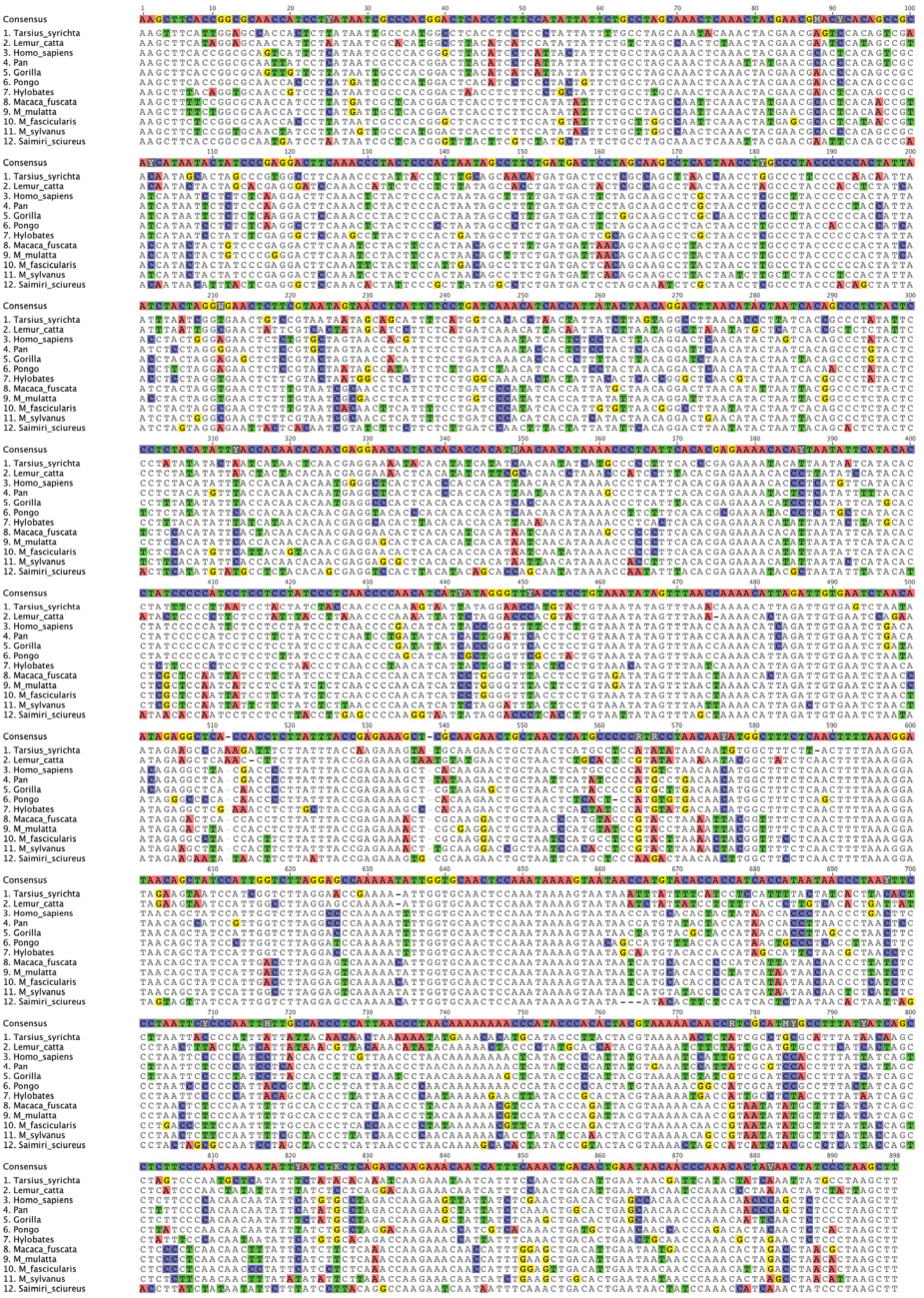

Figure 1.2 A small multiple sequence alignment of mitochondrial sequence fragments isolated from 12 species of primate. The alignment has 898 alignment columns and the individual sequences vary in length from 893 to 896 nucleotides long. Individual differences from the consensus sequence are highlighted. 373/898 (41.5%) sites are identical across all 12 species and the average pairwise identity is 75.7%. The data matrix size is 10 776 (898 × 12) with only 30 gap states. This represents a case in which obtaining an accurate multiple sequence alignment from unaligned sequences is quite easy and taking account of alignment uncertainty is probably unnecessary for most questions.

The statistical treatment of phylogenetics was made feasible by Felsenstein (1981), who described a computationally tractable approach to computing the probability of the sequence alignment given a phylogenetic tree and a model of molecular evolution, $\Pr\{D|T, \Omega\}$. This quantity is known as the *phylogenetic likelihood* of the tree and can be efficiently computed by the peeling algorithm (see Chapter 3). The statistical model of evolution that Felsenstein chose was a continuous-time Markov process (CTMP; see Section 3.1). A CTMP can be used to describe the evolution of a single nucleotide position in the genome, or for protein-coding sequences, either a codon position or the induced substitution process on the translated amino acids. By assuming that sites in a sequence alignment are evolving independently and identically a CTMP can be used to model the evolution of an entire multiple sequence alignment.

Although probabilistic modelling approaches to phylogenetics actually pre-date Sanger sequencing (Edwards and Cavalli-Sforza 1965), it was not until the last decade that probabilistic modelling became the dominant approach to phylogeny reconstruction (Felsenstein 2001). Part of that dominance has been due to the rise of Bayesian inference (Huelsenbeck et al. 2001), with its great flexibility in describing prior knowledge, its ability to be applied via the Metropolis–Hastings algorithm to complex highly parametric models, and the ease with which multiple sources of data can be integrated into a single analysis. The history of probabilistic models of molecular evolution and phylogenetics is a history of gradual refinement; a process of selection of those modelling variations that have the greatest utility in characterising the ever-growing empirical data. The utility of a new model has been evaluated either by how well it fits the data (formal model comparison or goodness-of-fit tests) or by the new questions that it allows a researcher to ask of the data.

1.2 Coalescent theory

When a gene tree has been estimated from individuals sampled from the same population, statistical properties of the tree can be used to learn about the population from which the sample was drawn. In particular the size of the population can be estimated using Kingman's *n-coalescent*, a stochastic model of gene genealogies described by Kingman (1982). *Coalescent theory* has developed greatly in the intervening decades and the resulting *genealogy-based population genetics* methods are routinely used to infer many fundamental parameters governing molecular evolution and population dynamics, including *effective population size* (Kuhner et al. 1995), rate of population growth or decline (Drummond et al. 2002; Kuhner et al. 1998), migration rates and population structure (Beerli and Felsenstein 1999, 2001; Ewing and Rodrigo 2006a; Ewing et al. 2004), recombination rates and reticulate ancestry (Bloomquist and Suchard 2010; Kuhner et al. 2000).

When the characteristic time scale of demographic fluctuations are comparable to the rate of accumulations of substitutions then past population dynamics are 'recorded' in the substitution patterns of molecular sequences.

The coalescent process is highly variable, so sampling multiple unlinked loci (Felsenstein 2006; Heled and Drummond 2008) or increasing the temporal spread

of sampling times (Seo et al. 2002) can both be used to increase the statistical power of coalescent-based methods and improve the precision of estimates of both population size and substitution rate (Seo et al. 2002).

In many situations the precise functional form of the population size history is unknown, and simple population growth functions may not adequately describe the population history of interest. Non-parametric coalescent methods provide greater flexibility by estimating the population size as a function of time directly from the sequence data and can be used for data exploration to guide the choice of parametric population models for further analysis. These methods first cut the time-tree into segments, then estimate the population size of each segment separately according to the coalescent intervals within it.

Recently there has been renewed interest in developing mathematical modelling approaches and computational tools for investigating the interface between population processes and species phylogenies. The *multispecies coalescent* is a model of gene coalescence within a species tree (Figure 1.3; see Section 2.5.1 and Chapter 8 for further details). There is currently a large amount of development of phylogenetic inference techniques based on the multispecies coalescent model (Bryant et al. 2012; Heled and Drummond 2010; Liu et al. 2008, 2009a,b). Its many advantages over standard phylogenetic approaches centre around the ability to take into account real differences in the underlying gene tree among genes sampled from the same set of

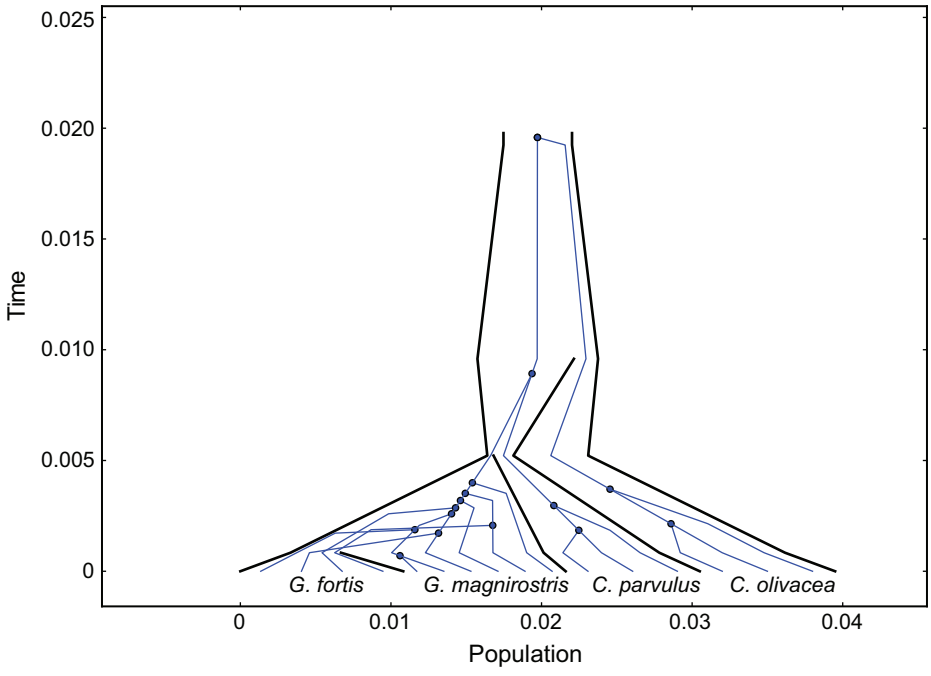

Figure 1.3 A four-taxa *species tree* with an embedded gene tree that relates multiple individuals from each of the sampled species. The species tree has (linear) population size functions associated with each branch, visually represented by the width of each species branch on the *x*-axis. The *y*-axis is a measure of time.

individuals from closely related species. Due to *incomplete lineage sorting* it is possible for unlinked genes from the same set of multispecies individuals to have different gene topologies, and for a particular gene to exhibit a *gene tree* that has different relationships among species than the true species tree. The multispecies coalescent can be employed to estimate the common species tree that best reconciles the coalescent-induced differences among genes, and provides more accurate estimates of divergence time and measures of topological uncertainty in the species tree. This exciting new field of coalescent-based species tree estimation is still in its infancy and there are many promising directions for development, including incorporation of population size changes (Heled and Drummond 2010), isolation with migration (Hey 2010), recombination and lateral gene transfer, among others.

1.3 Virus evolution and phylodynamics

A number of good recent reviews have been written about the impact of statistical phylogenetics and evolutionary analysis on the study of viral epidemiology (Kühnert et al. 2011; Pybus and Rambaut 2009; Volz et al. 2013). Although epidemic modelling of infectious diseases has a long history in both theoretical and empirical research, the term *phylodynamics* has a recent origin reflecting a move to integrate theory from mathematical epidemiology and statistical phylogenetics into a single predictive framework for studying viral evolutionary and epidemic dynamics. Many RNA viruses evolve so quickly that their evolution can be directly observed over months, years and decades (Drummond et al. 2003a). Figure 1.4 illustrates the effect that this has on the treatment of phylogenetic analysis.

Molecular phylogenetics has had a profound impact on the study of infectious diseases, particularly rapidly evolving infectious agents such as RNA viruses. It has given insight into the origins, evolutionary history, transmission routes and source populations of epidemic outbreaks and seasonal diseases. One of the key observations about rapidly evolving viruses is that the evolutionary and ecological processes occur on the same time scale (Pybus and Rambaut 2009). This is important for two reasons. First, it means that neutral genetic variation can track ecological processes and population dynamics, providing a record of past evolutionary events (e.g. genealogical relationships) and past ecological/population events (geographical spread and changes in population size and structure) that were not directly observed. Second, the concomitance of evolutionary and ecological processes leads to their interaction that, when non-trivial, necessitates joint analysis.

1.4 Before and beyond trees

Sequence alignment: After obtaining molecular sequences, a multiple sequence alignment must be performed before they can be analysed. Sequence alignment is a huge topic in itself (Durbin et al. 1998; Rosenberg 2009), and many techniques, including

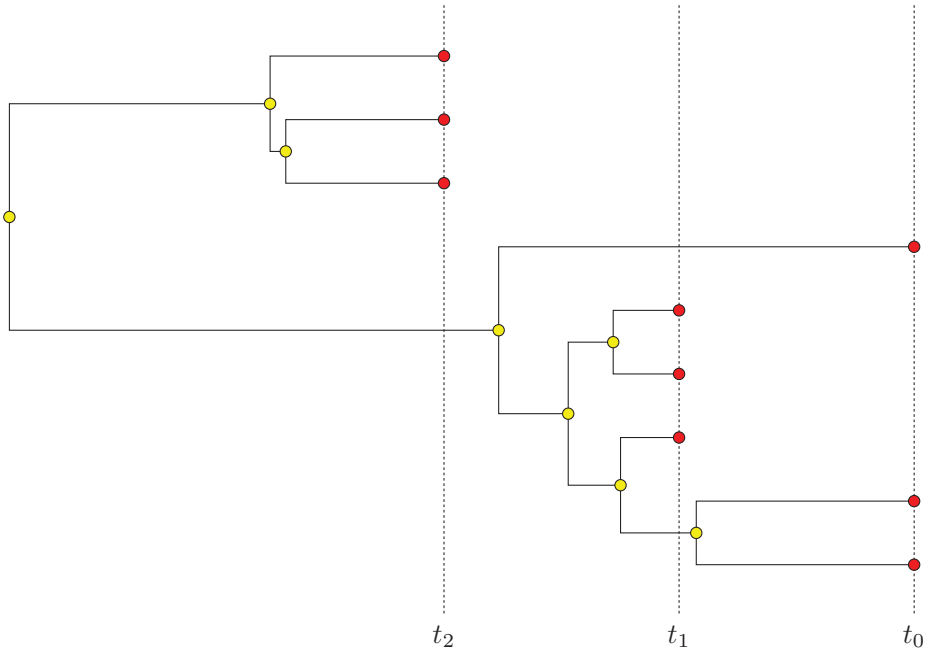

Figure 1.4 A hypothetical serially sampled gene tree of a rapidly evolving virus, showing that the sampling time interval ($\Delta t = t_2 - t_0$) represents a substantial fraction of the time back to the common ancestor. Red circles represent sampled viruses (three viruses sampled at each of three times) and yellow circles represent unsampled common ancestors.

dynamic programming, hidden Markov models and optimisation algorithms have been applied to this task. ClustalW and ClustalX (Larkin et al. 2007) are limited but widely used programs for this task. The Clustal algorithm uses a guide tree constructed by a distance-based method to progressively construct a multiple sequence alignment via pairwise alignments. Since most Bayesian phylogenetic analyses aim to reconstruct a tree, and the Clustal algorithm already assumes a tree to guide the alignment, it may be that a bias is introduced towards recovering the guide tree. This guide tree is based on a relatively simple model and it may contain errors resulting in sub-optimal alignments. T-Coffee (Notredame et al. 2000) is another popular program that builds a library of pairwise alignments to guide the construction of the complete alignment.

With larger data sets comes a need for high-throughput multiple sequence alignment algorithms; two popular and fast multiple sequence alignment algorithms are MUSCLE (Edgar 2004a,b) and MAFFT (Katoh and Standley 2013, 2014; Katoh et al. 2002)

However, a more principled approach in line with the philosophy of this book is to perform *statistical alignment* in which phylogenetic reconstructions and sequence alignments are simultaneously evaluated in a joint Bayesian analysis (Arunapuram et al. 2013; Bradley et al. 2009; Lunter et al. 2005; Novák et al. 2008; Redelings and Suchard 2005; Suchard and Redelings 2006). It has been shown that uncertainty in the alignment can lead to different conclusions (Wong et al. 2008), but in most cases it is

hard to justify the extra computational effort required, and statistical alignment is not yet available in BEAST 2.2.

Ancestral recombination graphs: A phylogenetic tree is not always sufficient to reflect the sometimes complex evolutionary origin of a set of *homologous* gene sequences when processes such as *recombination*, *reassortment*, *gene duplication* or *lateral gene transfer* are involved in the evolutionary history.

Coalescent theory has been extended to account for recombination due to homologous crossover (Hudson 1990) and the *ancestral recombination graph* (ARG) (Bloomquist and Suchard 2010; Griffiths and Marjoram 1996; Kuhner 2006; Kuhner et al. 2000) is the combinatorial object that replaces a phylogenetic tree as the description of the ancestral (evolutionary) history. However, in this book we will limit ourselves to trees.

1.5 Probability and Bayesian inference

At the heart of this book is the idea that much of our understanding about molecular evolution and phylogeny will come from a characterisation of the results of random or stochastic processes. The sources of this randomness are varied, including the vagaries of chance that drive processes like mutation, birth, death and migration. An appropriate approach to modelling data that are generated by random processes is to consider the probabilities of various hypotheses given the observed data. In fact the concept of probability and the use of probability calculus within statistical inference procedures is pervasive in this book. We will not, however, attempt to introduce the concept of probability or inference in any formal way. We suggest (Bolstad 2011; Brooks et al. 2010; Gelman et al. 2004; Jaynes 2003; MacKay 2003) for a more thorough introduction or reminder about this fundamental material. In this section we will just lay out some of the terms, concepts and standard relationships and give a brief introduction to Bayesian inference.

1.5.1 A little probability theory

A random variable X represents a quantity whose value is uncertain and described by a probability distribution over all possible values. The set of possible values is called the sample space, denoted $\mathcal{S}_X$.

A probability distribution $\Pr(\cdot)$ on discrete mutually exclusive values, x in sample space $\mathcal{S}_X$ (i.e. $x \in \mathcal{S}_X$), sums to 1 over all values so that:

$$\sum_{x \in \mathcal{S}_X} \Pr(x) = 1,$$

and $0 \leq \Pr(x) \leq 1$ for all x. In this case we say X is discrete.

A classic example of $\mathcal{S}_X$ is the set of faces of a dice, $\mathcal{S}_X = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$, and for a random variable representing the outcome of rolling a fair dice, $\Pr(X = x) = 1/6$ for all $x \in \mathcal{S}_X$.