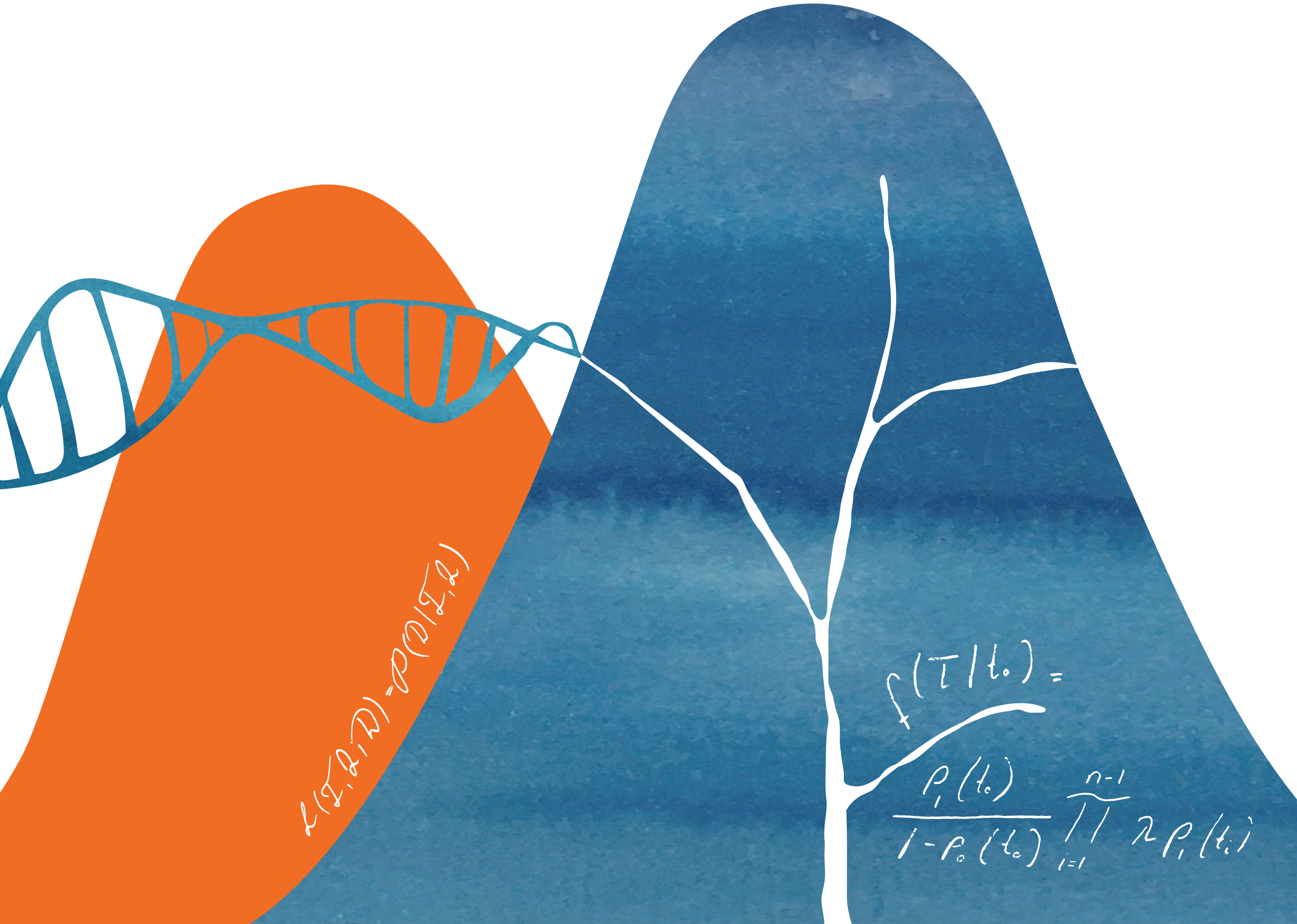
DECODING **GENOMES**

From **Sequences** to **Phylodynamics**

Tanja Stadler · Carsten Magnus · Timothy Vaughan
Joëlle Barido-Sottani · Veronika Bošková · Jana S. Huisman · Jūlija Pečerska

Illustrated by Cecilia Valenzuela Agüí · Edited by Jūlija Pečerska


$L(T,\theta,D) = P(D|T,\theta)$
$f(T|t_0) = \frac{p_1(t_0)}{1-p_0(t_0)} \prod_{i=1}^{n-1} \lambda p_1(t_i)$

DECODING GENOMES
From Sequences to Phylodynamics

Tanja Stadler,
Carsten Magnus, Timothy Vaughan

Joëlle Barido-Sottani, Veronika Bošková,
Jana S. Huisman, Jūlija Pečerska

Illustrated by Cecilia Valenzuela Agüí

Edited by Jūlija Pečerska

DECODING GENOMES: From Sequences to Phylodynamics

Tanja Stadler, Carsten Magnus, Timothy Vaughan, Joëlle Barido-Sottani, Veronika Bošková, Jana S. Huisman, Jūlija Pečerska.
Both the co-lecturers (C. Magnus, T. Vaughan) and the four teaching assistants (J. Barido-Sottani, V. Bošková, J. S. Huisman, J. Pečerska) are listed as authors in alphabetical order.

Illustrated by Cecilia Valenzuela Agüí. Edited by Jūlija Pečerska.

First edition, April 2024.

Additional information: https://decodinggenomes.org/.

Reporting errata: errata@decodinggenomes.org.

Feedback: feedback@decodinggenomes.org.

DOI: 10.3929/ethz-b-000664449

Preface

The journey leading to this book started in 2015 when we — the Computational Evolution (cEvo) group at ETH Zürich — designed a new Master course "Molecular evolution, phylogenetics, and phylodynamics" within the Computational Biology Master program. Our Master's students had undergraduate degrees in areas such as mathematics, computer science, physics, biology, or other life sciences. We did not identify a ready-to-use book catering to these different backgrounds. Tanja Stadler, Carsten Magnus, and Timothy Vaughan developed a new course building upon the existing wealth of literature. This course has been taught each year since 2015 at ETH Zürich.

The first generation of PhD students in the cEvo group (led by Tanja Stadler since 2014) were not only fantastic teaching assistants, but also offered to write down notes while we taught — producing a script for our students. This script was shared with the students in the years to come and refined through their critical comments. In 2017, we then decided to put the material into a book. Two generations of PhD students and one pandemic later, we are proud to present our book!

Thanks to everyone supporting us throughout. First, thanks to all our students for their critical feedback. Thanks to new cEvo group members who worked through our script upon starting research in the cEvo group and gave valuable feedback. A special thanks goes to Alexei Drummond, who taught some of the Bayesian material during his sabbatical visits and gave valuable feedback on our course content. Furthermore, we are immensely grateful to Oliver Pybus for inspiring discussions on phylodynamics throughout the years. The following researchers provided excellent feedback on different aspects of the book (in alphabetic order): Catharine Aquino, Dr. Jack Kuipers, Dr. Sophie Seidel, and Antoine Zwaans. And, finally, thanks to the whole cEvo group for providing a great atmosphere throughout the long and sometimes bumpy process of writing this book!

We invite you to visit `https://decodinggenomes.org/` for a PDF of the book, additional information, and a list of known errata. If you find any errors yourself, please report them to us via `errata@decodinggenomes.org`. You are welcome to send any other feedback to us via `feedback@decodinggenomes.org`.

Contents

Contents v

1 Introduction and Basics **1**
1.1 Overview . . . 1
1.1.1 Application areas . . . 3
1.1.2 Guide through the book . . . 5
1.2 Basics on evolution . . . 10
1.2.1 Lamarckian evolution . . . 10
1.2.2 Darwinian evolution . . . 10
1.2.3 Mendelian inheritance — foundations of genetics . . . 11
1.2.4 Genetic drift . . . 12
1.2.5 Modern synthesis (neo-Darwinism) . . . 13
1.2.6 Deoxyribonucleic acid (DNA) . . . 13
1.2.7 The central dogma of biology . . . 14
1.2.8 Exceptions to the central dogma . . . 16
1.2.9 Errors in replication . . . 17
1.2.10 Darwin today . . . 18
1.3 Basic definitions and concepts of probability . . . 19
1.3.1 Conditional probability . . . 21
1.3.2 Mathematical areas using probability relevant to this book . . . 22

2 Sequencing **27**
2.1 Overview . . . 27
2.2 First generation: Sanger sequencing . . . 29
2.2.1 Separation . . . 30
2.2.2 Amplification . . . 31
2.2.3 Sequencing . . . 31
2.2.4 Fluorescent labels . . . 32
2.3 Second generation: next-generation sequencing . . . 34
2.3.1 Pyrosequencing with 454 from Roche (discontinued in 2016) . . . 34
2.3.2 SOLiD from Thermo Fisher (discontinued) . . . 36
2.3.3 Ion Torrent from Thermo Fisher . . . 37
2.3.4 Illumina sequencers . . . 37
2.4 Third generation: Single-molecule sequencing . . . 39
2.4.1 PacBio RS from Pacific Biosciences . . . 41
2.4.2 MinION/GridION/PromethION from Oxford Nanopore . . . 42

3 Sequence alignments **45**
3.1 Pairwise alignments . . . 47
3.1.1 Dot-matrix method . . . 48
3.1.2 Scoring schemes . . . 49
3.1.3 Exhaustive method . . . 50
3.1.4 Dynamic algorithms . . . 51
3.1.5 Heuristic alignments: BLAST . . . 56
3.2 Multiple sequence alignments . . . 61
3.2.1 Heuristic multiple sequence alignment methods . . . 61
3.3 From sequencing reads to genome sequences . . . 64
3.3.1 Short-read alignment methods . . . 64
3.3.2 *De novo* assembly methods . . . 65

4 Genetic associations **69**
4.1 Testing for associations . . . 70
4.1.1 The case-control setup . . . 70
4.1.2 Calculating the p-value in a GWAS . . . 71
4.1.3 Correcting for multiple testing . . . 73
4.2 Potentials and drawbacks . . . 74

5 Molecular evolution **83**
5.1 Substitutions vs. mutations . . . 83
5.2 General theory on nucleotide substitution models . . . 84
5.2.1 Substitution rate matrix . . . 84
5.2.2 Transition probability matrix . . . 85
5.2.3 Stochastic process of sequence evolution . . . 94
5.2.4 Markov chain model of sequence evolution . . . 95
5.3 Common nucleotide substitution models . . . 99
5.3.1 JC69 model . . . 99
5.3.2 K80 model . . . 102
5.3.3 F81 model . . . 104
5.3.4 More general nucleotide substitution models . . . 105
5.3.5 Time scale: calendar time versus evolutionary time . . . 106
5.3.6 Time-reversibility of the nucleotide substitution models . . . 107
5.3.7 Site dependency in molecular substitution models . . . 111
5.4 Distance estimation for nucleotide sequences . . . 111
5.4.1 Simple pairwise distances . . . 112
5.4.2 Pairwise distances for JC69 using the method of moments . . . 113
5.4.3 Pairwise distances using a maximum likelihood approach for JC69 . . . 115
5.5 Allowing for rate variation across sites . . . 119
5.6 Amino acid substitution models . . . 121
5.6.1 Definition of amino acid substitution models . . . 124
5.6.2 JC69-like distance estimation for amino acid sequences . . . 125

5.7 Codon substitution models . . . 126
5.7.1 Definition of codon substitution models . . . 126
5.7.2 Detecting selection: d_N/d_S ratio . . . 128
5.7.3 Counting method . . . 129

6 Phylogenetic trees 135
6.1 Introduction to phylogenetic trees . . . 135
6.2 The mathematics of phylogenetic trees . . . 138
6.2.1 The mathematical definition of a phylogenetic tree . . . 138
6.2.2 The Newick tree format . . . 140
6.2.3 Counting trees . . . 140
6.3 Inferring phylogenies . . . 144
6.3.1 Phenetic approach: Distance-based methods . . . 146
6.3.2 Cladistic approach: Parsimony method . . . 155
6.3.3 Probabilistic approach: maximum likelihood methods . . . 164
6.4 From unrooted trees to time trees . . . 170
6.4.1 Time-reversibility implies that differently rooted trees have the same likelihood . . . 170
6.4.2 Rooting the tree . . . 172
6.4.3 Adding a calendar time scale to rooted phylogenetic trees . . . 172
6.5 Searching the tree space . . . 175
6.6 Examples of applications of phylogenetic reconstruction methods . . . 178
6.6.1 The first phylogenies . . . 178
6.6.2 Phylogenetics can reveal the origin of an emerging infectious diseases . 178
6.6.3 The HIV epidemic in Switzerland . . . 181
6.6.4 Phylogenetics in the court . . . 181

7 Statistical testing 185
7.1 Test whether to reject the null model $\mathcal{H}_0$ . . . 185
7.2 Test whether to reject the null model $\mathcal{H}_0$ in favour of $\mathcal{H}_1$ . . . 186
7.2.1 Likelihood ratio tests (LRT) . . . 186
7.2.2 Errors in statistical testing . . . 190
7.3 Compare models $\mathcal{H}_0, \mathcal{H}_1, \mathcal{H}_2, \ldots$ using the Akaike information criterion . . . 191
7.4 Assessing uncertainty in estimates . . . 193
7.4.1 Obtaining confidence intervals using the LRT . . . 193
7.4.2 Obtaining confidence intervals by redoing experiments . . . 195
7.4.3 Obtaining confidence intervals by non-parametric bootstrapping . . . 195
7.4.4 Obtaining confidence intervals by parametric bootstrapping . . . 196
7.4.5 Tree uncertainty estimation . . . 196

8 Traits and comparative methods 201
8.1 Assessing associations between discrete characters . . . 201
8.1.1 Assuming independence across samples . . . 202
8.1.2 Considering phylogenetic relatedness . . . 203

8.1.3 Detecting associations between discrete variables in empirical data . . . 205
8.2 Assessing associations between continuous characters . . . 205
8.2.1 Assuming independence across individuals . . . 206
8.2.2 Modelling continuous trait evolution with Brownian motion . . . 208
8.2.3 Considering phylogenetic relatedness using the contrast method . . . 208
8.2.4 Examples using the contrast method . . . 217
8.3 Extensions . . . 219
8.4 Connections to GWAS . . . 219

9 Phylodynamics 223
9.1 Birth-death models . . . 224
9.1.1 Population dynamic model . . . 224
9.1.2 Phylodynamic model . . . 231
9.1.3 Ranked labelled tree topologies . . . 234
9.1.4 Expected population sizes and branching times . . . 236
9.1.5 Distribution of branching times . . . 243
9.1.6 Applications . . . 250
9.2 Coalescent theory . . . 256
9.2.1 The Wright-Fisher process . . . 256
9.2.2 Kingman's coalescent process . . . 259
9.2.3 Population dynamics . . . 263
9.2.4 Coalescent approximation of birth-death models . . . 267
9.2.5 Effective population size . . . 269
9.2.6 Application . . . 269
9.3 Comparison of coalescent and birth-death models . . . 272
9.4 Phylodynamics for non-binary trees with co-occurring internal nodes . . . 273
9.5 Accounting for population structure . . . 274
9.5.1 Population structure shapes phylogenies . . . 275
9.5.2 Structured birth-death phylodynamic models . . . 277
9.5.3 Structured coalescent phylodynamic models . . . 284
9.5.4 Related structured models . . . 296

10 Bayesian inference 301
10.1 Bayesian theory . . . 301
10.1.1 Maximum likelihood vs. Bayesian inference . . . 303
10.1.2 Example: Bayesian inference of pairwise genetic distance . . . 304
10.1.3 Priors . . . 305
10.1.4 Incorporating additional data into a Bayesian analysis . . . 306
10.1.5 Reporting uncertainty: credible intervals . . . 307
10.1.6 Bayesian model selection . . . 308
10.1.7 A Bayesian phylogenetics and phylodynamics framework . . . 309
10.2 Markov chain Monte-Carlo (MCMC) sampling . . . 311
10.2.1 MCMC algorithm . . . 313
10.2.2 Toy example . . . 316

10.2.3 Mixing rate and effective sample size 317
10.2.4 MCMC for phylogenetics and phylodynamics 319
10.2.5 Time complexity of phylogenetic MCMC algorithms 320
10.3 Applications 321
10.3.1 Epidemiology: The Ebola epidemic in West Africa, 2013-2016 322
10.3.2 Epidemiology: The SARS-CoV-2 pandemic 322
10.3.3 Macroevolution: Phylogeny of penguins 323

11 Phylogenetic networks **327**
11.1 Mathematics of phylogenetic networks 327
11.2 Recombination networks 328
11.2.1 Homologous recombination processes 329
11.2.2 Impact of recombination on phylogenies 331
11.2.3 Wright-Fisher model with recombination 333
11.2.4 Coalescent with recombination 334
11.2.5 Bayesian inference 335
11.2.6 Applications 337
11.3 Species networks 341

12 Opportunities and challenges **345**
12.1 Opportunities through novel applications 345
12.1.1 Applications in biology and the life sciences 345
12.1.2 Applications in anthropology 351
12.2 Statistical and computational challenges 352
12.2.1 Analysing large datasets 352
12.2.2 Combining genetic sequences with additional data 353
12.2.3 Determining model adequacy 355
12.3 Final words 356

List of Figures **359**

List of Tables **365**

List of Boxes **367**

Index **369**

Bibliography **379**

1 Introduction and Basics

Nothing in biology makes sense except in the light of evolution.
(Theodosius Dobzhansky (1973))

1.1 Overview

In biology, we study organisms to understand how the living world functions. However, we cannot directly observe and measure every aspect of the living world. Some features are not observable because we have not (yet) developed the proper technical equipment, while other features may be unobservable *per se*. For example, many species populate this planet, but we cannot directly observe how they came about as this process occurred millions of years before any of us were born. In the domain of epidemiology, we can observe which human hosts are infected by a pathogen, but we cannot directly observe the dynamics of the pathogen infecting a new human host, as we would aim to prevent the infection instead of watching it. Statistical inference methods can help us understand such unobservable processes using available snapshot data, such as data collected from extant species or infected hosts.

This book focuses on statistical and computational methods for learning about unobservable evolutionary and population dynamic processes using genome data (data on the genetic material carried by individuals), namely deoxyribonucleic/ribonucleic acid (DNA/RNA) sequence data, possibly together with some phenotypic data (data on the appearance of individuals). *Evolution* refers to the change of populations through time with respect to the heritable features of the individuals that make up those populations. Heritable features are, in particular, the genomes of individuals and their phenotypic features. *Population dynamics* refers to the change of populations in size and density through time and across space.

A classic area of biology where statistical and computational tools are required to understand an unobservable process is macroevolution, where the biological unit is a species, and the available data include both genotypic and phenotypic information. Phylogenetics (Chapters 6 to 8 and 11 of the book) was initially introduced to study the evolutionary relationships between species and has been used more recently to study relationships between other biological units such as infected individuals or single cells.

The central object in phylogenetics is a *phylogeny*, which may be a *tree* or a *network*. A phylogeny starts with one individual, and its offspring are tracked through time. In a tree, each

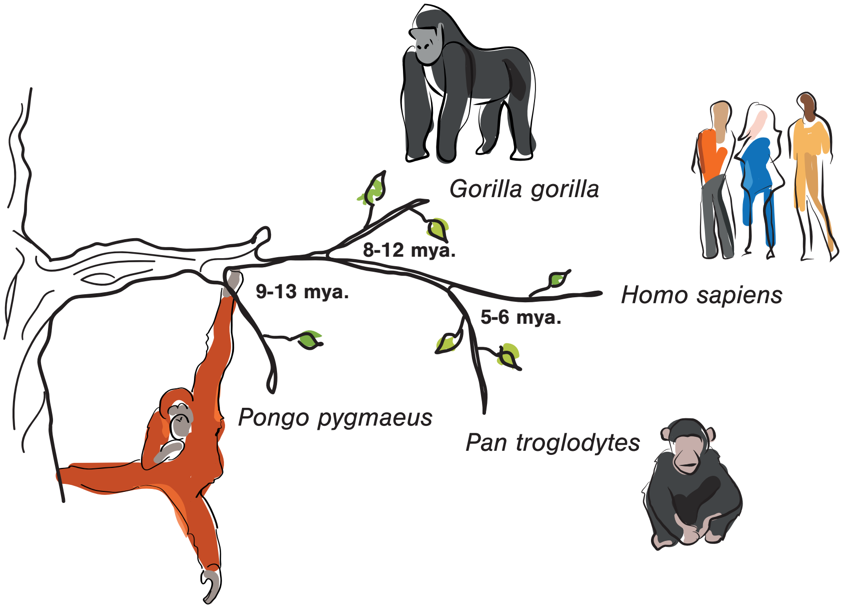


Figure 1.1: Phylogeny of the great apes, consisting of humans (*Homo sapiens*) and their closest relatives. The phylogenetic tree shows that humans and chimpanzees (*Pan troglodytes*) diverged around 5-6 million years ago from a common ancestor, while the gorilla (*Gorilla gorilla*) is a more distant relative, diverging from the common ancestral lineage around 8-12 million years ago. The orangutans (*Pongo pygmaeus*) diverged from the ancestral lineage already 9-13 million years ago.

offspring has precisely one parent, while in a network, offspring may have one or more parents. Phylogenetic methods aim to reconstruct the phylogeny based on genomes and possibly phenotypic features of the sampled individuals, such as the present-day species. Figure 1.1 shows a phylogenetic tree of great apes inferred using phylogenetic methods.

Taking the analysis one step further, phylodynamics (Chapters 9 and 10 of the book) aims to describe and quantify the population dynamic processes that gave rise to the phylogeny. In macroevolution, the main population dynamic processes are speciation and extinction: how quickly species appear and go extinct. A further macroevolutionary process is hybridisation, where two or more species give rise to an offspring species. Both phylogenetics and phylodynamics use models of molecular evolution (Chapter 5 of the book) to capture the way the genomes of individuals of some biological unit, for example, a species, change through time.

The book is structured into four main parts, each containing examples of real-world data analysis results obtained using the computational and statistical tools presented.

Obtaining and organising sequences: how do we obtain sequences from biological samples, align them, and what can data mining tell us about them? (Chapters 2 to 4)

Molecular evolution: how does genetic information change through time? (Chapter 5)

Phylogenetics: how can we determine the relatedness of biological samples based on their genetic information? What is their underlying phylogeny? How do phenotypes evolve along a phylogeny? (Chapters 6 to 8, 10 and 11)

Phylodynamics: what population dynamics (e.g. speciation and extinction or pathogen transmission dynamics) give rise to the phylogeny and the genetic and phenotypic information we observe? (Chapters 9 and 10)

We end with a chapter on applications of the presented statistical and computational methodology across biological domains and a discussion of ongoing methodological challenges (Chapter 12).

The remainder of the introduction is structured as follows. We first briefly outline areas where phylogenetics and phylodynamics are used. Next, we give a detailed overview of the content and structure of this book (Section 1.1.2). We then provide the basics on evolution, the process leading to changes in the genomes and phenotypic features, and thus, a core principle behind phylogenetics and phylodynamics (Section 1.2). Finally, we end the introduction with basic definitions and concepts of probabilities used throughout the book (Section 1.3).

1.1.1 Application areas

“Nothing in biology makes sense except in the light of evolution,” the title of an essay by evolutionary biologist Theodosius Dobzhansky (1973), highlights that we must acknowledge evolutionary processes when studying any area of biology. It follows that phylogenetics and phylodynamics are crucial to understanding unobservable processes in a wide range of biological areas beyond evolutionary biology and even non-biological areas, including, but not limited to, fields listed below.

Macroevolution, where the biological unit is a species;

- Molecular evolution describes genetic changes in the species;
- Phylogeny displays the relationship between species, that is, the species tree or network;
- Population dynamics describes speciation and extinction.

Microevolution, where the biological unit is a uni- or multicellular individual (such as a bacterial or archaeal cell, a unicellular eukaryote, or a multicellular individual);

- Molecular evolution describes genetic changes in individuals;
- Phylogeny displays the relationship between individuals;
- Population dynamics describes the birth and death of individuals.

Infectious disease epidemiology, where the biological unit is an infected host;

- Molecular evolution describes genetic changes in the pathogen population within an infected host and bottlenecks at transmission;
- Phylogeny displays the pathogen transmission chain;
- Population dynamics describes the transmission of the pathogen to susceptible hosts and the recovery or death of infected hosts.

Immunology, where the biological unit is an immune response cell within a host, such as a B- or T-cell;

- Molecular evolution describes changes in immune cells through, for example, somatic hypermutations or recombination;
- Phylogeny displays the immune cells' evolutionary relationship;
- Population dynamics describes the generation and loss of different immune cells within the host.

Development, where the biological unit is a cell within an organism;

- Molecular evolution rarely happens in somatic cells due to very good repair mechanisms at cell division; however, genetic barcodes that mutate fast and thus undergo molecular evolution during an experiment can be inserted into cells (Wagner and Klein 2020);
- Phylogeny displays the relationships of different cells;
- Population dynamics describes the division, differentiation, and death mechanisms of different cells (different cells form via cell differentiation from stem cells).

Cancer, where the biological unit is a cell within an organism;

- Molecular evolution describes the genetic changes of the cells;
- Phylogeny displays the relationships of different cancer cells and healthy cells;
- Population dynamics describes the emergence, spread, and loss of different cancer cells.

Linguistics, where the anthropological unit (rather than biological unit) is a language;

- Evolution (which is not molecular here) describes the changes in words and grammatical structures within languages through time;
- Phylogeny displays the evolutionary relationships of languages;
- Population dynamics describes the appearance and extinction of languages.

1.1.2 Guide through the book

This book aims to provide readers coming from different backgrounds (including mathematics, statistics, biology, computer science, and physics) with an understanding of the kind of information encoded in genetic sequences (to answer questions such as “Why do genetic sequence analysis?”). Furthermore, the reader will acquire the necessary skills to understand, plan, and perform genetic sequence analysis using data mining, phylogenetic, and phylodynamic techniques (“How to do an analysis?”). Throughout, we provide examples (“What can be learnt from such an analysis?”). We anticipate the needs of readers with different backgrounds by explaining fundamental concepts from biology and mathematics in the form of boxes. Moreover, at the end of the introduction, we provide a short section on evolution and a short section on probabilities.

In the remainder of this section, we overview this book’s specific content and structure. The book covers the steps from obtaining genetic sequence data from DNA to performing a comprehensive data analysis. The first step in a genetic sequence analysis is to obtain the sequence, that is, to transform the genetic information encoded by the individual of interest to a format that we can use and analyse. In particular, we want to represent the individual’s DNA as a sequence of the letters `A`, `C`, `G`, and `T`. To do so, we need to extract the DNA from cells and then use sequencing techniques to read and decipher this DNA. In the case of RNA viruses, RNA is extracted from virions, reverse-transcribed into DNA, and then sequenced. The book, therefore, begins with an introduction to genetic sequencing technologies (Chapter 2). For the purpose of this book, we assume that these sequences fully characterise the individuals in our subject population, and we do not consider epigenetic patterns.

The next step is to *align* the sequences from the sequenced individual to one another. In an alignment, different sequences are typically displayed in different rows, and their nucleotides are assigned to columns or sites such that the nucleotides across individuals of one site evolved from a single ancestor. Differences in these sites across individuals mean genotypic variation and may determine phenotypic variation. The differences contain information regarding evolutionary history and evolutionary and population dynamic processes. In Chapter 3, we discuss different methods for aligning sequences. We discuss methods for obtaining alignments optimally and explain the basic idea of *heuristic approaches* for alignments, that is, fast approaches that do not guarantee optimality. In particular, we introduce BLAST (`https://blast.ncbi.nlm.nih.gov/Blast.cgi`) (Basic Local Alignment Search Tool, Altschul et al. (1990) and Altschul et al. (1997)), which lets us find the homologues (see Chapter 3) to a single sequence by comparing it against a huge library of sequences. The BLAST algorithm is the first *data mining approach* (an approach aimed at finding associations within a large dataset) discussed in this book and is one of the most widely used algorithms in bioinformatics.

In fact, sequencing technologies often do not return the whole sequence representing an individual (in particular if the genome of the individual is very large), but many sequence fragments, which are called *reads*. The reads need to be joined to obtain the whole sequence. This procedure is called *assembly*. After assembly, an alignment of the sequenced individuals can

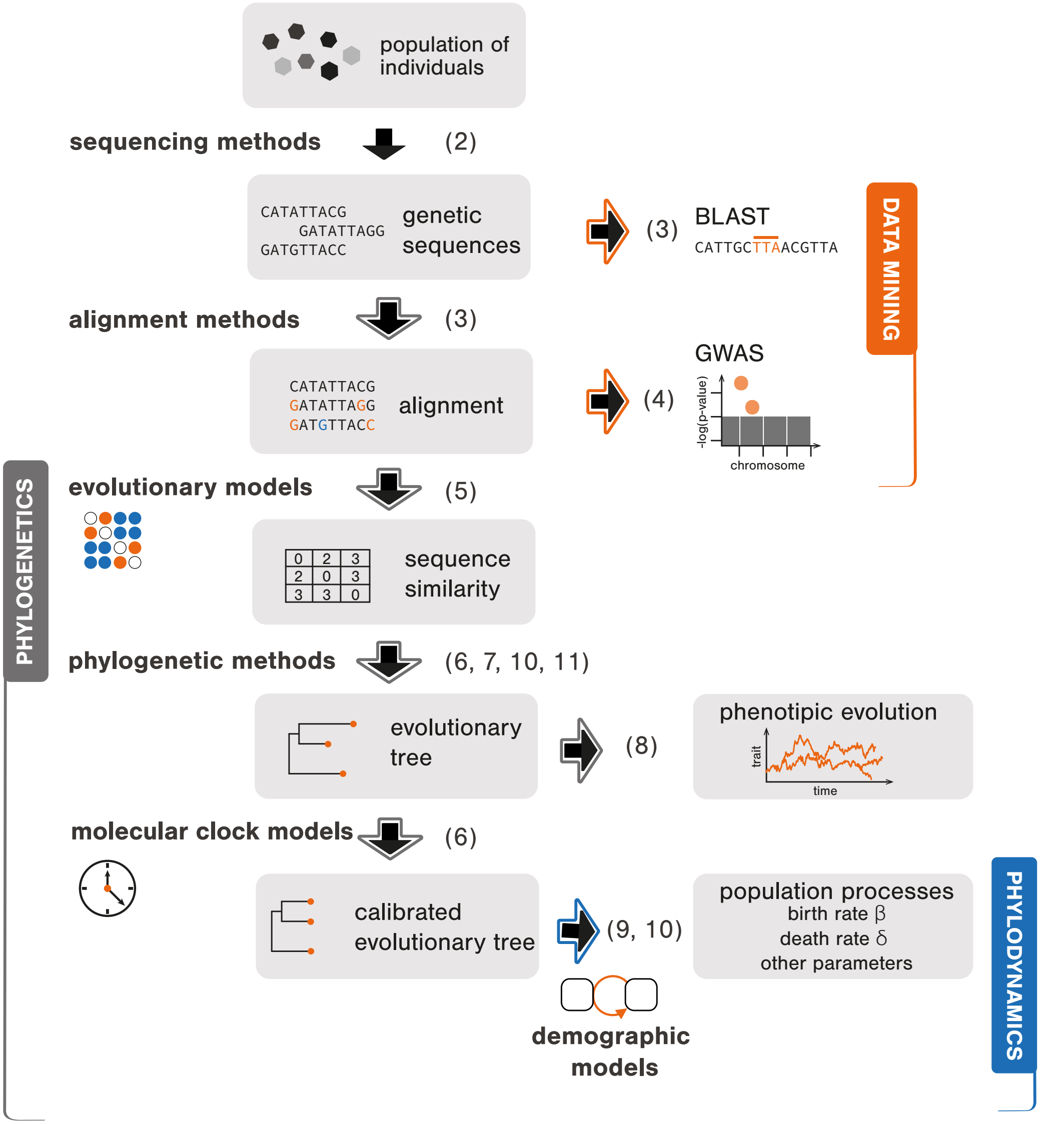


Figure 1.2: A guide through the book: from sequences to phylodynamics. The numbers in parentheses indicate the chapters in which the topics are explained.

be obtained. Assembly procedures rely on alignment methods and thus are also discussed in Chapter 3.

The alignment encodes information we aim to extract to answer biological questions. We continue with another data mining approach, namely *genome-wide association studies, (GWAS)* (Chapter 4). In GWAS, the aligned genome sequences obtained from multiple individuals are considered together with traits of these individuals (e.g. increased risk of a certain type of cancer) to detect if certain genome variants or mutations are associated with those traits. We highlight that whole books have been written on such data mining approaches (e.g. Aggarwal (2015)), and we only provide the main ideas here.

Data mining approaches such as GWAS assume each site in an alignment is an independent data point. This assumption is valid for genome data stemming from different human individuals since recombination quickly breaks up the linkage between the sites of interest. However, if genomes accumulate substitutions along a phylogenetic tree (meaning there are no recombination or other non-tree processes), then the individuals close in the phylogeny share more similarities (such as identical nucleotides at a site) than distantly related individuals. Thus, the sites are not independent data points, and GWAS approaches are not appropriate. Joseph Felsenstein (1985b) explicitly spelt out the need to acknowledge the phylogeny when analysing genotypes and their association with certain traits (Chapter 8).

In Chapters 6 to 10, we consider data where sites are linked due to a shared evolutionary history. The shared evolutionary history can be displayed in a phylogeny. The overarching aim is to reconstruct this phylogeny based on the sequences and then to infer the evolutionary and population dynamic processes giving rise to the phylogenies.

In Figure 1.2, the left bracket groups the aspects of sequence analysis that fall in the field of *phylogenetics*, the study of evolutionary history. In phylogenetics, we assume a (molecular) evolution model (Chapter 5) and then reconstruct phylogenies (Chapters 6 and 10) based on the sequence or morphological data. Based on the phylogenies, we can further investigate processes occurring along the phylogenetic tree. In particular, we can obtain an understanding of how genes and genotypes change along the phylogeny through time, that is, understand molecular evolution processes (Chapter 7). Furthermore, we can assess how traits change along the phylogeny, with the phylogeny representing genotypic change (Chapter 8). Such analyses shed light on phenotypic evolution processes. In particular, as in a GWAS, the relationship between the genotype and the phenotype is addressed, now with statistical tools acknowledging dependencies between sites. Importantly, all molecular or phenotypic evolutionary processes are assumed not to influence the tree; instead, they occur on a given (unknown) tree. In turn, the sequence data from the evolutionary processes are used to reconstruct this tree.

The field of phylogenetics goes back to 1837 when Charles Darwin sketched a phylogenetic tree in his notebook, shown in Figure 1.3. However, the computational birth of this field only occurred in 1957, when Michener and Sokal published a paper on a computational tool that allows reconstructing a phylogenetic tree from sequence data (Michener and Sokal 1957). This first tool was based on the simple principle that similarity between two individuals indicates

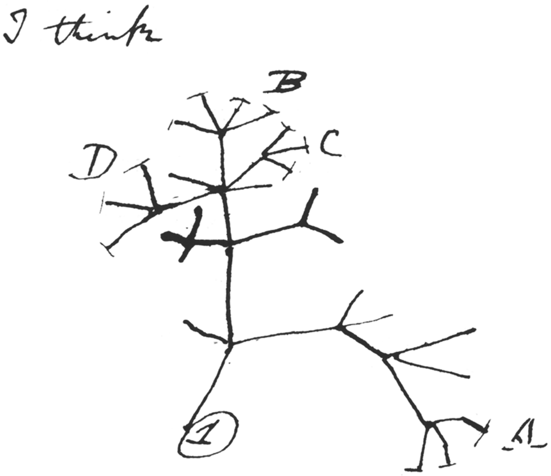


Figure 1.3: A first sketch of an evolutionary tree by Charles Darwin, from his 1837 notebook (Darwin 1837).

that they are close relatives, whereas dissimilarity indicates a more distant relationship. Joseph Felsenstein revolutionised phylogenetic tree inference in the 1980s by introducing statistical tools allowing the maximum likelihood and Bayesian approaches to be applied to phylogenetics (Felsenstein 1981). Initially, phylogenetics was developed and used in macroevolution. Later, starting with the studies on HIV in particular, phylogenies of viruses were reconstructed to understand their evolution and epidemiology.

The field of *phylodynamics*, denoted in Figure 1.2 by the bottom right bracket, studies how processes give rise to and shape phylogenies. Phylodynamic approaches fit population dynamic models (e.g. models of speciation/extinction, models of pathogen transmission, and so on) to the reconstructed phylogenies. In particular, these approaches take into account that the phenotype may influence the shape of the phylogeny. This allows us, for example, to quantify fitness differences across individuals and thus to quantify selective advantages of certain phenotypes or assess the global migration pattern of individuals. Phylodynamic methods require a time tree — a tree with branches in units of calendar time. Charles Darwin sketched such a time tree (Figure 1.4). The first key papers on phylodynamics of macroevolution appeared in the 1990s (Nee, May and Harvey 1994; Harvey, May and Nee 1994). However, the field only started flourishing after the publication by Grenfell et al. (2004), considering the phylodynamics of pathogens. Phylodynamics is discussed in Chapters 9 and 10.

Approaches presented in this book thus far assume that there is no linkage across sites (Chapter 4) or that the sites evolved along a phylogenetic tree, meaning there is complete linkage across sites of an alignment (Chapters 6 to 10). However, there is increasing evidence that evolutionary histories are best modelled by a “Network of Life”, meaning an individual

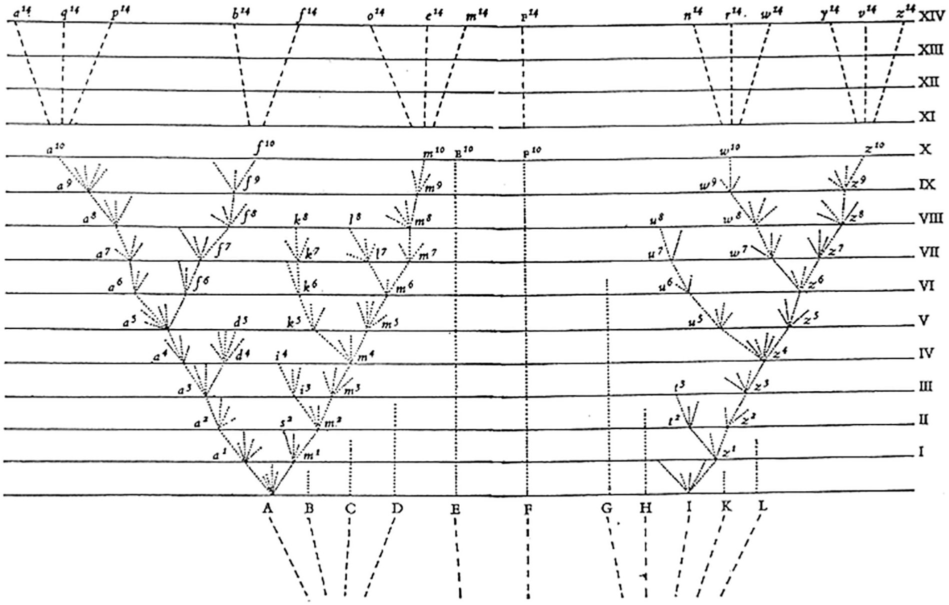

a^{14} q^{14} p^{14} b^{14} f^{14} o^{14} e^{14} m^{14} F^{14} n^{14} r^{14} w^{14} y^{14} v^{14} z^{14}
XIV
XIII
XII
XI
a^{10} f^{10} m^{10} E^{10} F^{10} w^{10} z^{10}
X
a^{9} f^{9} m^{9} w^{9} z^{9}
IX
a^{8} f^{8} k^{8} l^{8} m^{8} u^{8} w^{8} z^{8}
VIII
a^{7} f^{7} k^{7} l^{7} m^{7} u^{7} w^{7} z^{7}
VII
a^{6} f^{6} k^{6} m^{6} u^{6} z^{6}
VI
a^{5} d^{5} k^{5} m^{5} u^{5} z^{5}
V
a^{4} d^{4} i^{4} m^{4} z^{4}
IV
a^{3} i^{3} m^{3} t^{3} z^{3}
III
a^{2} s^{2} m^{2} t^{2} z^{2}
II
a^{1} m^{1} z^{1}
I
A B C D E F G H I K L

Figure 1.4: Sketch of a time tree with horizontal lines drawn every 1000 generations. This is the only figure in the "Origin of Species" by Charles Darwin (1859).

has two or more parents as a result of reticulate evolution (examples are hybridisation, horizontal gene transfer, or recombination), while in a tree each individual can only have one parent. Reticulate evolution causes some sites to have different evolutionary histories (due to the different parents). In other words, some sites are not linked. Importantly, some sites are also linked, making GWAS approaches unsuitable.

We end the methodological part of the book by presenting basic concepts regarding *phylogenetic networks* (Huson, Rupp and Scornavacca 2010) which are required in intermediate scenarios between phylogenetic trees — where sites are completely linked, — and GWAS — where all sites are unlinked (Chapter 11).

Throughout the book, we provide empirical examples of the introduced methodology, mainly from the classic fields in phylogenetics and phylodynamics, namely macroevolution and virus epidemiology. In the final chapter, we outline applications of the tools presented throughout the book in fields where phylogenetics and phylodynamics are entering now (see also Section 1.1.1), together with methodological challenges we need to overcome to make full use of the data.

1.2 Basics on evolution

Evolution, the process that gives rise to changes in genotypes in populations through time, is at the core of all the statistical and computational methods discussed in this book. Thus, in this section, we will briefly describe the main aspects of the theory of evolution and how this theory itself evolved over time.

Throughout human history, people have been trying to explain how the living world came into existence. The idea that everything was created at once and has since existed in a fixed state was eventually replaced by the concept of evolution and perpetual change. Extant species data and fossil evidence provide overwhelming support for evolution.

Initially, the concept of evolution was discarded by many, as it contradicted religious views of men being the pride of creation. Nowadays, the concept of evolution is widely accepted in the scientific community and significant parts of society. Some parts of society continue to reject this concept in favour of creationism (see Matzke (2016) for an interesting view on the evolution of creationism). Evolution as a scientific theory also went through its own stages of evolution, from simple beginnings to current, more elaborate concepts. This section introduces theories that paved the way to the current understanding. It is by no means a complete picture of the evolution of evolution, but it focuses on how we came to our current understanding of evolution through some specific influential historical concepts.

1.2.1 Lamarckian evolution

In the nineteenth century, biologists and naturalists increasingly discussed the possibility of explaining species diversity via evolution. In 1809, the French biologist Jean-Baptiste Lamarck proposed that evolution occurs through the use and disuse of features (Lamarck 1809). This means that an organism could develop a useful feature during its lifetime, which would then be passed on to its offspring. This is the definition of soft inheritance, the inheritance of acquired characteristics.

Lamarck's favourite example was the giraffe, shown in Figure 1.5. He explained the length of the giraffe's neck as follows: the first giraffes had short necks that made it hard for them to reach the leaves on the trees. This meant that giraffes always had to stretch their necks, which would become slightly lengthened over the course of their lives. Their offspring would then inherit this lengthened neck, and, over the course of many generations, the neck length would increase to its current proportion. Thus, evolution occurs via individuals locally adapting to the environment and their offspring inheriting the acquired characteristics.

1.2.2 Darwinian evolution

In 1859, the British biologist Charles Darwin published his book "The Origin of Species by Means of Natural Selection" (Darwin 1859) describing a theory of evolution that aims to

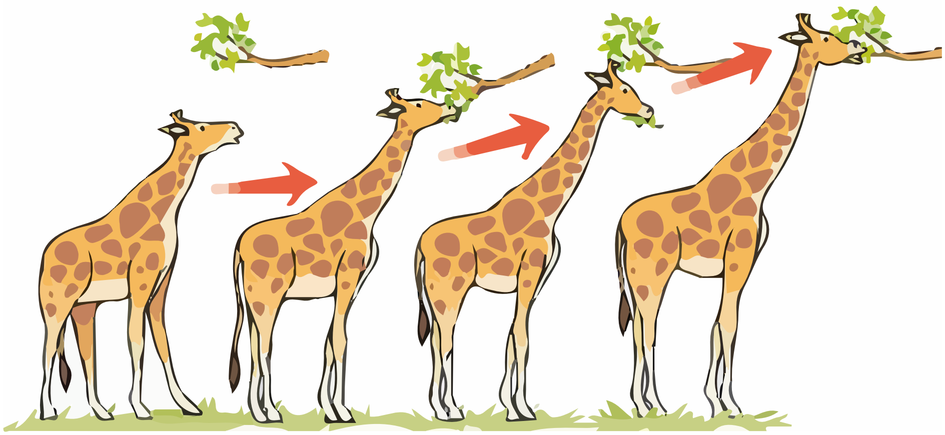


Figure 1.5: Example of Lamarckian evolution. According to Lamarck, evolution occurs through the use and disuse of features. Thus, if giraffes prefer leaves from taller branches, they will strain their necks more and more during their lifetime, which would cause their offspring to have longer necks (Lamarck 1809).

explain evolution via the mechanism of natural selection. For evolution to occur via natural selection, four components are needed:

(i) **Multiplication**: the individuals multiply and produce offspring;

(ii) **Variation**: there is phenotypic variation across individuals, that is, individuals differ in some aspect of their appearance;

(iii) **Heredity**: the phenotype is to some extent heritable from one generation to the next;

(iv) **Competition**: there are fitness differences across phenotypes, meaning that the average number of surviving offspring depends on the phenotype of the parent individual.

While Darwin's mechanism could explain the data he collected during his voyage on the Beagle, one important feature was missing in the theory of natural selection: how phenotypes are inherited could not be explained thoroughly.

1.2.3 Mendelian inheritance — foundations of genetics

In 1866, seven years after Darwin's influential book appeared, the Austrian monk Gregor Mendel published his observations on the possible mechanisms of heredity of the phenotypes, which he developed from his experiments with pea plants (Mendel 1866). He observed that certain traits, such as flower colours, get passed on to the next generation in a predictable fashion. He described invisible factors — which we now call *genes*, — that have different variants — which we now call *alleles*. The variants determine the traits — which we call *phenotypes*. The invisible factors are passed on from one generation to the next. We introduce here vocabulary that is important throughout this book.

Gene: hereditary unit encoding a protein, which, in turn, defines (part of) a trait (e.g. the pea flower colour);

Allele: the version of a gene (e.g. one allele of the colour gene may encode for white (y); another allele for purple (Y));

Genotype: the collection of genes of one individual;

Phenotype: the collection of traits of one individual.

Based on the experimental data, Mendel concluded that each pea plant has two alleles of each gene, a random one from the father and a random one from the mother. A *dominant allele* is the allele that determines the phenotype (e.g. if the purple allele Y is dominant, peas having the Yy allele combination will bloom purple); the other, *recessive allele*, is overruled by the dominant allele and will only have an effect if both the alleles inherited from the parents are recessive.

Weismann (1893) later performed experiments showing that only the genes in the germ line (cell line that produces gametes or sex cells in sexually reproducing organisms) are passed on to the next generation, and no changes in other cells of the organism (somatic cells) are passed on.

This concept of inheritance of genes from the germ line is a core concept in modern biology. It opposes the concept of Lamarckian evolution, which suggested that acquired phenotypes were inherited. Nowadays, Gregor Mendel is acknowledged as the founder of the field of genetics and as the person who closed the gap in Darwin's work by explaining the inheritance mechanism. However, Mendel's work was widely ignored for at least 30-40 years after publication before the connection to Darwin's theory was made.

1.2.4 Genetic drift

Apart from evolving due to natural selection, populations may also evolve due to pure chance. There is chance involved in which individuals reproduce and which alleles they pass to their offspring. When considering a small population, it is evident that these chance processes can lead some alleles to be lost or take over the population. This mechanism of "genetic drift" for evolution was introduced by Sewall Wright (1955). The work of Motoo Kimura (1968) emphasised the importance of such a "neutral" evolutionary process — compared to selection processes.

1.2.5 Modern synthesis (neo-Darwinism)

The term “modern synthesis” was coined by Julian Huxley in his 1942 book “Evolution: The Modern Synthesis” (Huxley 1942). This theory reflects the consensus theory of evolution, which combines Darwin’s theory of evolution through natural selection, neutral evolution, gene flow between populations, and Mendelian genetics, with the latter providing the mechanism for inheritance.

Evolution may occur through natural selection acting on the phenotypes. The phenotypes are encoded through the genotype, and the genotype in the germ line is the unit that is passed on to offspring following the rules of Mendelian genetics. In parallel, evolution may occur through genetic drift (pure chance); again, the inheritance follows the rules of Mendelian genetics. Finally, gene flow between populations can also impact evolution (Andrews 2010).

Many people contributed to the establishment of this theory. Theodosius Dobzhansky, as well as Rosemary and Peter Grant, demonstrated that the modern synthesis theory holds up if one tests it in natural populations (Dobzhansky 1937; Grant et al. 1976). George Simpson showed that paleontological data, which is evidence for the process of evolution in the past, is in accordance with the modern synthesis theory (Simpson 1944). By now, modern synthesis is the primary accepted mechanism of evolution (but see Shapiro and Noble (2021) and Rose and Oakley (2007) for shortcomings with respect to evolutionary mechanisms such as epigenetics and horizontal gene transfer).

This book introduces neutral sequence evolution models in Section 5.3. In Sections 9.1 and 9.2, we consider neutral evolution models at the population level. We take into account selection in the sequence evolution models in Sections 5.5 to 5.7 and model selection at the population level in Section 9.5.

1.2.6 Deoxyribonucleic acid (DNA)

The next step to further our understanding of evolution was to understand how genes and genotypes are encoded.

In 1871, Friedrich Miescher published work on isolating and identifying the deoxyribonucleic acid, or DNA (Miescher-Rüsch 1871), which encodes the genotype. Together with images by Rosalind Franklin, this allowed us to decipher the molecular structure of DNA — a double helix made of two linked strands — in 1953 (Watson and Crick 1953).

The double-stranded DNA helix has the same structure in all known biological entities (eukaryotes, bacteria, archaea, and viruses) on Earth, and it consists of a sugar-phosphate backbone to which nitrogenous bases, namely purine or pyrimidine bases, are attached. The sugar, the phosphate, and the nitrogenous base make up a single nucleotide — a DNA building block, the deoxynucleotide triphosphate (dNTP).

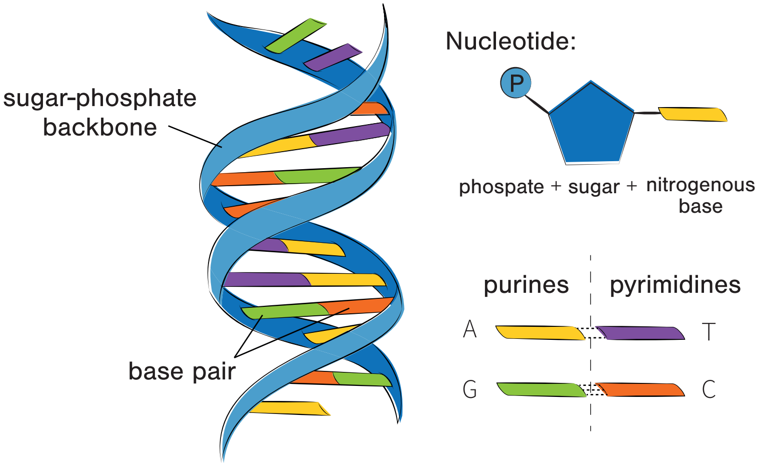


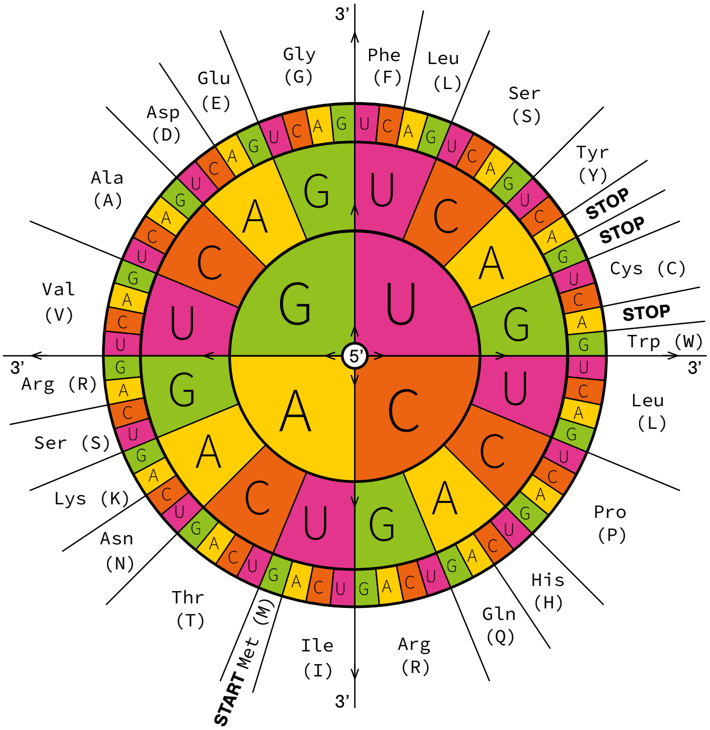
Figure 1.6: DNA double helix displaying how the nucleotides are arranged into a double-stranded molecule (left). The structure of each nucleotide — a sugar-phosphate backbone with a nitrogenous base — and their grouping into purines and pyrimidines is displayed on the right.

We refer to a nucleotide by the name of its nitrogenous base. There are four nucleotides, two purines, *adenine* (A) and *guanine* (G), and two pyrimidines, *cytosine* (C) and *thymine* (T). The successive order of these four nucleotides determines an individual's genotype. A few viruses are RNA-based (ribonucleic acid) rather than DNA-based. The RNA strand is composed of the same A, C, G, but uses the pyrimidine *uracil* (U) instead of T. Again, the successive order of these four nucleotides is the genotype of the virus.

We call the order of nucleotides a *genetic sequence*. In the double-stranded helix, one strand is complementary to the other. The nucleotides are paired in the helix according to strict compatibility rules such that A on one strand is complemented by T on the other strand, whereas G is always complemented by C (see Figure 1.6). This means that if we know the sequence of one strand, we can always determine the complementary sequence of the other strand, and thus, only one strand is reported in genetic sequence data. The analysis of genetic sequence data is the main focus of this book.

1.2.7 The central dogma of biology

The central dogma of modern biology describes the flow of genetic information within a biological system. It states that information flows from DNA to RNA through *transcription* and from RNA to proteins through *translation*, without any informational exchange flowing back (Figure 1.8 illustrates this principle flow of information).



Amino acid	3-letter code	1-letter code
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Figure 1.7: The codon sun shows the encoding of amino acids by triplets of nucleotides. The nucleotide at the first position of the triplet is chosen from the innermost circle; the second and third are then picked from the second and the third circles from the centre, respectively. The amino acids encoded by the nucleotide triplets are displayed on the outermost circle as three- and one-letter codes. The table on the right shows amino acid names and their corresponding codes. Thus, for example, the triplet `TCG` encodes the amino acid Serine (`Ser`, `S`).

The part of the genotype that encodes proteins is referred to as genes. Each group of three successive nucleotides in genes is called a *codon*. Of the $4^3 = 64$ possible codons, three codons terminate RNA translation. The remaining 61 codons translate into the 20 amino acids, meaning several codons encode for the same amino acid. Amino acids are abbreviated with one- or three-letter codes. Proteins are characterised by the sequence of amino acids. Figure 1.7 depicts which codon translates into which amino acid.

The genome also consists of non-coding regions, where the nucleotides do not code for a gene. Instead, the non-coding regions may serve regulatory functions, or may not have a function, or have some yet unknown function. Despite not producing proteins directly, we highlight that these regions may play a role in determining the phenotype (the organism's appearance)

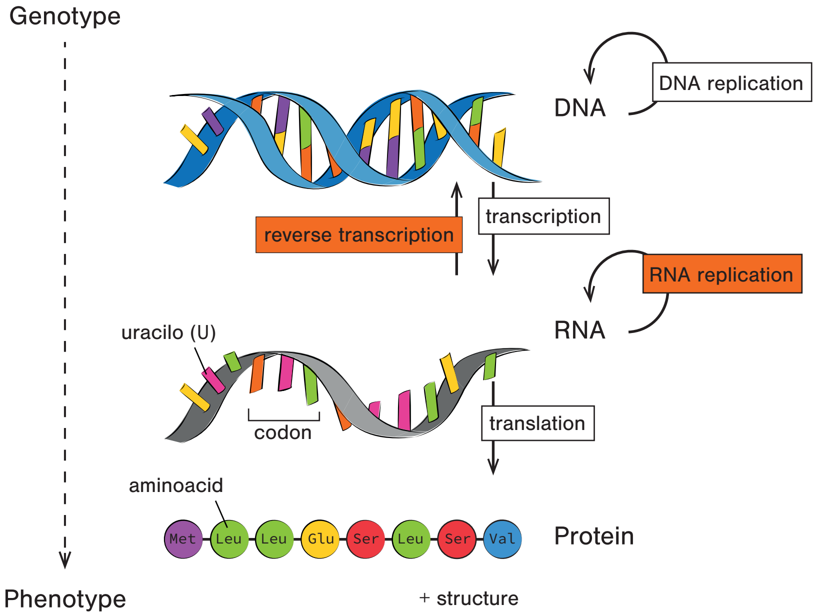


Figure 1.8: The central dogma of biology states that information flows from DNA to RNA (transcription) to proteins (translation). An exception to this dogma is reverse transcription, where virus RNA is reverse-transcribed into DNA. Further, not only DNA but also RNA may replicate.

regardless.

In summary, according to the central dogma, information flows from DNA (the genotype) to RNA to proteins (the phenotype), meaning the genotype determines the phenotype (but see Section 1.2.8 for an exception). According to Darwin, natural selection acts on that phenotype.

The famous evolutionary biologist Richard Dawkins proposed an analogy to baking, in which he compared the genotype to the recipe and the phenotype to the cake (Dawkins 2009). This statement nicely points out a particularity of this connection — the recipe is not necessarily bad just because we failed to make the cake once. In the context of cell biology, this means that even if there was an error during a single run of transcription or translation, next time, the transcription and translation might be perfectly normal again, yielding the expected protein.

1.2.8 Exceptions to the central dogma

Generally, there is no rule without an exception, and this is also the case for the central dogma of biology. In fact, *reverse transcription* of RNA to DNA is possible. Figure 1.8 shows this exception with an upward arrow.

Reverse transcription (Baltimore 1970) was discovered[1] in a specific class of viruses, the so-called *retroviruses*. For example, the human immunodeficiency virus (HIV) is a retrovirus. It stores all its genetic information as RNA and transfers its genetic material into the host cell with a reverse transcriptase enzyme, which reverse-transcribes the RNA into DNA. The produced DNA is then incorporated into the host cell's genetic material by the HIV integrase enzyme and is transcribed and replicated using standard host cell machinery.

We note that RNA viruses may also replicate without the reverse transcription to DNA by copying the RNA directly. In fact, this kind of replication is performed by many well-known viruses, such as influenza, Ebola, and corona- and polioviruses.

1.2.9 Errors in replication

The variation of genotypes between cells — and, by extension, between organisms — arises due to the error-prone replication of DNA during cell division. Errors may be introduced into the copied DNA strand during DNA replication when a cell prepares for division, for example, due to polymerase error or external mutagens (such as chemicals, UV radiation, and so on). The following errors can happen during DNA replication, resulting in a different sequence of nucleotides in the copied strand compared to the template strand.

Point mutation: during the production of the DNA copy, a wrong nucleotide is built into the sequence with respect to the template, producing a copy where a single character is replaced by another (this character is said to have *mutated*). We refer to a point mutation as a *mutation* throughout this book;

Recombination: the combination of "parental" genetic sequences to produce a "child" sequence. For example, in eucaryotic cells, a block of nucleotides is exchanged between chromosome pairs;

Insertion and deletion: extra nucleotides are inserted into the copy or lost compared to the template. Such events are referred to as *indels* (*in*sertions and *del*etions);

Repeats and inverted repeats: in a repeat, a sequence of k nucleotides in the template strand is inserted several times into the copied DNA strand. We obtain an inverted repeat if a reverse copy of length k occurs in the copied DNA.

A detailed understanding of these mechanisms is not crucial for our purposes. However, it is crucial for this book to note that errors can occur during replication. These errors cause variation in genotypes, which in turn produces variation in phenotypes.

[1] This discovery was a component of the research for which David Baltimore, Renato Dulbecco, and Howard Martin Temin received the 1975 Nobel Prize in Physiology or Medicine.

1.2.10 Darwin today

Due to errors in replication, different individuals of the same population can have slightly different genotypes and, thus, phenotypes. Selection can then act on this phenotype, leading to the propagation of certain variants in the population and the extinction of other variants.

To illustrate this, let us look at the interaction between a host's immune system and a virus population, particularly how the immune system puts selective pressure on the virus. The cells of the immune system use the proteins on the surface of the *virions* (virus particles) to identify them as foreign entities and attack them. The immune system tries to eliminate the virions, and by doing so, it exerts selective pressure on the virus population. The virions have an evolutionary advantage if they have a variant of the surface protein that the host's immune system does not recognise. This viral variant is linked to a mutation in the viral genotype; the virions possessing this mutation will propagate, while the virions without this mutation will be eliminated. After some time, the entire virus population will have the mutated genotype: evolution at the genotypic level driven by selection at the phenotypic level.

As mentioned above, evolution can also be the result of pure chance or genotypes flowing from the outside into a population.

By combining our knowledge from different scientific disciplines (such as genetics and molecular evolution) into modern theory and using Darwin's four components in addition to neutral drift and gene flow processes, we can summarise as follows.

(i) **Multiplication**: DNA replicates. Somatic cell replication produces more cells within an individual; germ line cell replication may give rise to an offspring individual. In either case, the genotypes determine the phenotypes.

(ii) **Variation**: variation in the phenotype across individuals is observed due to mutations, recombination, insertions and deletions (indels), as well as (possibly inverted) repeats in the genotype at replication.

(iii) **Heredity**: heredity of the phenotype occurs due to the passage of DNA (or RNA in the case of RNA viruses) from parent to offspring (via the germ line in multi-cellular organisms); the genotype of the offspring encodes its phenotype.

(iv) **Competition**: there are fitness differences across phenotypes, meaning the average number of surviving offspring depends on the phenotype of the parent individual.

(v) **Drift**: genotypes can expand or vanish purely by chance (genetic drift), meaning the average number of surviving offspring is determined by chance if only drift occurs.

(vi) **Gene flow**: populations' genotype composition may change from one generation to the next through gene flow from the outside.

Importantly, points (i) to (iv) result in evolution through natural selection, and points (i) to (iii) and (v) in neutral evolution.

This view completely ignores any possible impact of the environment on the phenotype. However, increasing evidence exists for the impact of the environment on the phenotype. The field of epigenetics studies mechanisms that change phenotypes beyond DNA modification. In particular, the activation pattern of genes and changes in this pattern were found to be such an epigenetic mechanism. Activation patterns vary due to the variation in molecules binding to DNA, such as a methyl group binding to DNA (DNA methylation). The activation patterns may change throughout an individual's lifetime due to environmental factors and, in fact, can also be inherited through the germ line. Such inheritance of environmentally acquired phenotypes brings us back to Lamarck.

One example of epigenetic effects is the phenotypic differences between identical twins. Genetically, these two individuals are identical (although there might be slight differences due to somatic mutations during development); however, the two individuals might not look exactly the same due to epigenetic differences acquired throughout the twins' lifetime (Fraga et al. 2005). Another example is a population of clonal bacteria — bacteria with the same genotype — in which some bacteria are more virulent than others, for example, groups of identical *Salmonella* bacteria that split into those that are more adept at entering epithelial cells in the gut, those that release toxins, and those that reproduce quickly. Their DNA methylation patterns are different due to environmental effects (Casadesús and Low 2006).

1.3 Basic definitions and concepts of probability

Consider a random experiment where the set of all possible outcomes is a discrete set called the *state space* Ω. A *probability* is defined as a function, denoted with a capital P, that

(i) maps each outcome in Ω to a number between 0 and 1 (including 0 and 1); this number is referred to as the probability of the outcome, and

(ii) the sum of the probabilities of all outcomes in the state space is 1.

A simple example is rolling a six-sided die one time. The possible outcomes are 1, 2, 3, 4, 5, or 6. Thus, the state space is $\Omega = \{1, 2, 3, 4, 5, 6\}$. In our example, the outcome "Rolling 1 when throwing a die once" is denoted by $\{1\}$. If the die is fair, each outcome is equally likely, and the probability of rolling a 1 is $1/6$. Formally, this is denoted by

$$P(\{1\}) = 1/6. \tag{1.1}$$

A more interesting example is rolling the fair die twice. There are 36 different possible outcomes, the state space is $\Omega = \{(1,1), (1,2), \ldots, (1,6), (2,1), \ldots, (6,6)\}$, and every single result has the probability $1/36$. If we are interested in the sum of the two numbers, we can calculate the probability of obtaining this sum by counting all possible results that lead to that number and dividing it by all possible results:

$$P(\text{sum} = 5) = P(\{(1,4),(2,3),(3,2),(4,1)\}) = {}^{4}\!/_{36}. \tag{1.2}$$

A function applied to the possible outcomes, such as summing the two numbers in the example above, is called a *random variable* and is denoted with a capital letter, for example, X. Note that a random variable is neither random nor a variable but a function. For a discrete set of outcomes, the associated random variable is said to be discrete. In our example, $X((i,j)) = i + j$ is a discrete random variable. For convenience, $P(X = x)$ is an abbreviation for "the probability of the set of outcomes for which we obtain x when we apply the function X to the possible outcomes of the random experiment". The value x is also called a realisation of X. In the example with two dice, this means

$$P(X = x) = P(\{(i,j) \in \{(1,1),\ldots,(6,6)\} \text{ for which } i + j = x\}). \tag{1.3}$$

The probability that some random variable X takes some value x is written as $P(X = x)$. However, when there is no ambiguity regarding the random variable the probability refers to, it is common to write $P(x)$ instead. Similarly, when there is no ambiguity with respect to the value, $P(X)$ may also be used.

A *stochastic model* formalises a random experiment (such as a die roll) to make predictions about this random experiment. It contains the state space Ω, the probability function P, and the random variable X. Note that all these entities were specified above for the die-rolling example, meaning we formulated a stochastic model for die-rolling.

There are also situations where the state space is continuous. A classic example is a person's height. The probability that a person is exactly 1.83m tall is 0. But the probability that a person's height ranges between 1.83 and 1.84m is > 0. In the case of a continuous random experiment, one denotes its *probability density* instead of the probability of each individual outcome. To obtain the probability of a continuous random variable taking the values in a specific set, one has to integrate the probability density over this set. For example, we denote height with the continuous random variable Y and the probability density of Y with f_Y. The probability that a person's height ranges between 1.83 and 1.84m then is

$$P(1.83 < Y < 1.84) = \int_{1.83}^{1.84} f_Y(x)\,\mathrm{d}x. \tag{1.4}$$

Two important measures are commonly reported for every distribution. These are the *mean* and the *variance*. The mean is the average value one would expect from a series of random experiments. For a random variable X, the mean is denoted by $\mathrm{E}(X)$. If the random variable is discrete, that is, has a discrete state space Ω, the mean is defined as the sum of all possible results weighted by the probability of obtaining these results:

$$\mathrm{E}(X) = \sum_{x \in \Omega} x P(X = x). \tag{1.5}$$

If the random variable is continuous, that is, it has a continuous state space Ω, and has a probability density f_X, the mean is defined as the integral of all possible values weighted by its probability density:

$$\mathrm{E}(X) = \int_{\Omega} x f_X(x) \, \mathrm{d}x. \tag{1.6}$$

Note that the mean is not necessarily a value that one can obtain as a result of the random experiment. For example, the mean outcome when throwing a fair six-sided die is 3.5, but that value is not included in the state space.

Var denotes the variance and is the average deviation from the mean, meaning that it measures how dispersed the distribution is. For a discrete random variable X with state space Ω, the variance is defined as

$$\mathrm{Var}(X) = \mathrm{E}((X - \mathrm{E}(X))^2) = \sum_{x \in \Omega} (x - \mathrm{E}(X))^2 P(X = x). \tag{1.7}$$

Similarly, the variance of a continuous random variable is defined as

$$\mathrm{Var}(X) = \mathrm{E}((X - \mathrm{E}(X))^2) = \int_{\Omega} (x - \mathrm{E}(X))^2 f_X(x) \, \mathrm{d}x. \tag{1.8}$$

The standard deviation is sometimes reported instead of the variance. *Standard deviation*, denoted by $\sigma(X)$, is the square root of the variance:

$$\sigma(X) = \sqrt{\mathrm{Var}(X)}. \tag{1.9}$$

For two random variables, the *covariance* is defined as

$$\mathrm{Cov}(X, Y) = \mathrm{E}\left((X - \mathrm{E}(X))(Y - \mathrm{E}(Y))\right). \tag{1.10}$$

This definition generalises the variance, as it holds that $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$.

1.3.1 Conditional probability

The concept of conditional probability is easiest to understand for discrete random variables. Consider a random experiment with discrete state space Ω and two sets of outcomes $A, B \subset \Omega$, with $P(B) \neq 0$. We denote the intersection of sets A and B with $A \cap B$. The conditional probability $P(A|B)$ is the probability that event A happens, given that we know B happened.

It can be calculated as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{1.11}$$

$P(A \cap B)$ is also called the joint probability, the probability of both A and B happening. Intuitively, we can understand that the above formula holds by considering $P(A \cap B) = P(A|B)P(B)$. Indeed, we can calculate the joint probability by first evaluating the probability of event B happening and then evaluating the probability of event A happening when knowing that B happened. This corollary,

$$P(A \cap B) = P(A|B)P(B), \tag{1.12}$$

is sometimes known as the product rule for probabilities.

Returning to our die experiment from above, imagine we want to determine the probability of scoring at least a sum of 10 when rolling the die twice (event A) while knowing that a doublet has been rolled (event B). We could only get a score of at least 10 with doublets when rolling $(5,5)$ or $(6,6)$. This means the probability of scoring a sum of ≥ 10 given that we threw a doublet is

$$P(A|B) = {}^{2}\!/_{6} = {}^{1}\!/_{3}. \tag{1.13}$$

Alternatively, we can get this probability by considering $P(A \cap B)$ and $P(B)$. Doublets are $B = \{(1,1),(2,2),(3,3),(4,4),(5,5),(6,6)\}$ and thus $P(B) = {}^{6}\!/_{36}$. Furthermore, we have $A \cap B = \{(5,5),(6,6)\}$ and thus $P(A \cap B) = {}^{2}\!/_{36}$, Now, we apply the right-hand side of the conditional probability equation above:

$$\frac{P(A \cap B)}{P(B)} = \frac{{}^{2}\!/_{36}}{{}^{6}\!/_{36}} = {}^{1}\!/_{3}. \tag{1.14}$$

When looking at continuous random variables, the concept of conditional probabilities follows the same ideas but becomes formally more difficult. We refer the interested reader to a textbook on probability theory, such as the book by Williams (2001).

1.3.2 Mathematical areas using probability relevant to this book

The concept of probability is used in different areas of mathematics. In this book, we will specifically encounter the following fields.

Probability theory studies the rules to calculate probabilities of events and random variables given that the underlying distribution is known. For example, we will derive the probability of sequences evolving along a phylogenetic tree (see Section 6.3.3).

Stochastic processes study successively repeated random experiments and the overall behaviour of the realisations of these processes. Here, we assume that the underlying probability distribution is known. We will cover Markov chains in Box 24 on page 98 and Brownian motion in Box 30 on page 209 as examples of stochastic processes. The stochastic processes are, for example, used to model changes in genotypes and phenotypes (see Chapters 5 and 8).

Statistics uses a set of observations to try and deduct information about the underlying distribution.

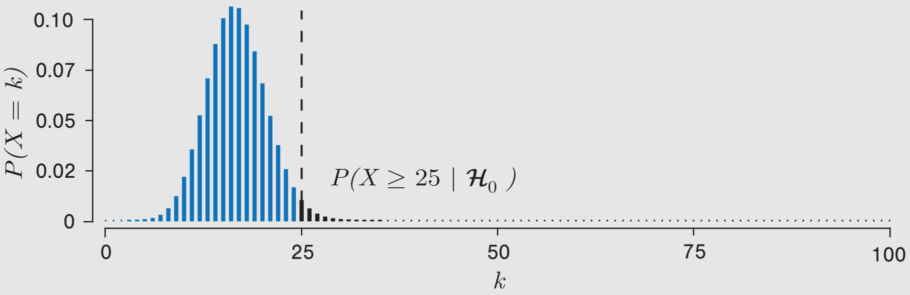
- In *parameter estimation*, we use observations to estimate a parameter of an *a priori* specified distribution, such as the probability of observing a particular outcome. For example, we estimate the speciation and extinction rates based on a phylogeny (see Chapter 9).
- In *hypothesis testing*, we use the observations to test whether the data support a specific hypothesis — formalised as a specific stochastic model giving rise to a probability distribution. Given such a hypothesis, we can calculate how likely it is to obtain the observed or more extreme outcomes. This probability is called the *p-value* (see Box 1 on page 24). The p-values will be employed to test whether the given data (the outcome of a probabilistic experiment) evolved under the assumed stochastic model. This value will be crucial in Chapters 4 and 7.
- In *uncertainty quantification*, the uncertainty of outcomes is quantified. We will focus on uncertainty in evolutionary parameter and tree estimation in Chapter 7.

Box 1: p-value

Definition: As introduced in Section 1.3, we use capital X to represent a random variable and small x to represent its realisation. Let us assume a *null hypothesis* $\mathcal{H}_0$, which, in mathematical terms, corresponds to some statement about the potential probability distribution of X. Given the null hypothesis is true, the probability of the outcome being x or more extreme is called the *p-value* for x under the null hypothesis. If more extreme means greater than x, the p-value is defined as $P(X \geq x|\mathcal{H}_0)$. If more extreme means smaller than x, the p-value is defined as $P(X \leq x|\mathcal{H}_0)$.

Example: For example, consider rolling a six-sided die $n = 100$ times. Let the random variable X be the number of sixes obtained. The null hypothesis $\mathcal{H}_0$ is that the six-sided die is fair, meaning that the probability that we roll a six in a single throw is $1/6$. Suppose we obtained $x = 25$ as a result of our experiment, and we want to determine the p-value.

The probability to obtain k sixes out of n independent die rollings is $\binom{n}{k}\frac{1}{6}^k\frac{5}{6}^{n-k}$, that is, X follows a *binomial distribution* with parameter p equal to $1/6$ (Box 3 on page 25; for a definition of the binomial coefficient $\binom{n}{k}$ see Box 2 on page 25). The distribution is displayed for $n = 100$ in the graph below. The p-value for our realisation ($x = 25$) is $p = P(X \geq 25|\mathcal{H}_0) = \sum_{k=25}^{100}\binom{100}{k}\frac{1}{6}^k\frac{5}{6}^{100-k} = 0.022$ (assuming more extreme is greater). The blue-coloured areas are the events at least as extreme as our outcome ($x = 25$).



Usage: One can pre-define a *significance level*, commonly denoted as α, which is typically set to $\alpha = 0.05$ or $\alpha = 0.01$. If a test is said to have a significance level α, the cumulative probability of a false positive is α. If the p-value for an obtained outcome is below α, the null hypothesis is rejected at the level α and is said to differ *significantly* from the null hypothesis. We can also say that such an observation is in the tail of the distribution.

We may also reject the null hypothesis if the outcome is in the left or right tail of the distribution. The significance level is then divided by two, and the hypothesis is rejected if $P(X \geq x|\mathcal{H}_0) < \alpha/2$ or $P(X \leq x|\mathcal{H}_0) < \alpha/2$. α is again the significance level, and $\alpha/2$ is referred to as *rejection threshold*. For further information on the p-values, see, for example, Dorey (2010).

Box 2: Binomial coefficient

The *binomial coefficient* is defined as

$$\binom{a}{b} = \frac{a!}{b!(a-b)!} \tag{B2.1}$$

for integers $a \geq b \geq 0$.

The expression $a!$ is called factorial and for $a > 0$ defined as

$$a! = a \times (a-1) \times (a-2) \times \ldots \times 2 \times 1. \tag{B2.2}$$

Further, 0! is defined to be 1.

The binomial coefficient counts the number of unordered subsets of size b, given a set of size a. Equivalently, the binomial coefficient counts the number of ways in which $a+b$ balls can be ordered by colour, where a balls are red and b black.

It is sometimes useful to generalise the definition of the binomial coefficient so that

$$\binom{a}{b} = 0 \text{ for } b > a. \tag{B2.3}$$

Box 3: Binomial distribution

Let m be a number of experiments with two possible outcomes: "success" (probability p) or "failure" (probability $1-p$). Each such experiment is called a *Bernoulli trial*. The random variable "number of successes among m experiments", X, follows the *binomial distribution* which has the probability function

$$P(X = k) = \binom{k}{m} p^k (1-p)^{m-k}. \tag{B3.1}$$

2 Sequencing

2.1 Overview

Genetic information serves as the blueprint to build an organism. This genetic information is contained in each cell of the organism in the form of *deoxyribonucleic acid (DNA)*. The language of these blueprints is universal for all living organisms on Earth: each individual is characterised by a particular order of DNA building blocks, the nucleotides. Moreover, this language is still the same for viruses that are not considered to be alive but also contain genetic information in the form of DNA or *ribonucleic acid (RNA)* within their virions. A *genetic sequence* is a digital (human- and computer-readable) excerpt from the blueprint, representing the order of nucleotides — adenine (A), cytosine (C), guanine (G), thymine (T) — as they appear in the DNA molecule (with uracil (U) instead of thymine (T) in the case of RNA). In other words, a sequence is a string composed of letters A, C, G, T/U. The sequence may represent the whole genome of the individual or just parts of the genome, such as single genes. *Sequencing* is the process of obtaining a genetic sequence from a biological sample. Box 4 on page 28 provides a short vocabulary to aid in understanding the sequencing techniques.

This chapter is a short introduction to DNA sequencing to give the reader an idea of the available platforms and their advantages and disadvantages. RNA is not directly sequenced as it is not as stable as DNA, and RNA molecules are instead reverse-transcribed to DNA before sequencing.

We provide an overview of the steps needed to prepare a biological sample for sequencing and then discuss DNA sequencing platforms that are currently in use or have been widely used in the past. We distinguish three generations of sequencing technologies. No single platform can be labelled the best, as each platform fulfils different specific needs, with respect to, for example, sequence length and coverage, and each has its drawbacks.

To start sequencing, we first need to acquire the genetic material we want to sequence or read. Cells are taken from an organism or a specific tissue of the organism, and the genetic material is isolated from the cells. DNA isolation requires breaking the cell membranes and molecules within the cell (also called lysing the cell) and then separating the non-DNA fragments from the DNA molecules. Commercial kits employing chemical methods are available to break up the cells, and then a centrifuge can be used to separate the DNA molecules from the non-DNA fragments. Isolating RNA (e.g. from the virions of RNA viruses) requires analogous steps. However, the procedure is generally more complicated for RNA, as the RNA molecule

Box 4: Sequencing glossary

Template: a stretch of DNA in the biological sample which will be sequenced;

Primer: a short DNA fragment complementary to (one end of) the template. The primer is essential for starting DNA polymerisation and, thus, the sequencing reaction;

Library: a mixture of different templates ready for sequencing;

Sequence: the order of building blocks (nucleotides) in a template or gene or genome;

Sequencing: the act of determining the order of nucleotides in the template;

Sequencing run: one round of operation of the sequencing machine;

Read: a single instance of output from a single sequencing run — it is a sequence representing a partial or whole template, often still containing errors;

Sequencing error: the difference between the sequence retrieved via sequencing and the template sequence;

Assembly: the process of combining the information in individual reads into the sequence of the genome.

is less stable and more prone to degradation. Once the RNA is isolated, it is directly reverse-transcribed to complementary DNA (cDNA) to ensure as little loss of information as possible due to RNA degradation.

Second, if the DNA/cDNA is too long for sequencing, we need to selectively amplify parts of the DNA/cDNA (e.g. using *polymerase chain reaction (PCR)*, see Figure 2.1 and Box 5 on page 29) or cut the DNA/cDNA into smaller pieces. Mechanical methods, such as sonification, break up the DNA in random places, or alternatively, restriction enzymes (stemming from bacteria) can be used to cut the DNA in specific locations. The DNA/cDNA fragments that will be sequenced are called templates.

To start the sequencing requires that the templates are physically separated and possibly amplified. Separation depends in detail on the particular sequencing technology and will be discussed in more detail as the individual technologies are introduced. Further, the first and second generations of sequencing methods presented below have in common that the signal from a single template instance would be too weak to be captured by the sequencer's detection technology. Thus, an *amplification* step is needed to intensify the signal, where the number of each template present is multiplied using PCR.

Sequencing technologies then provide us with the sequence of nucleotides in the templates. Many of these technologies produce only short reads or fragments of the biological sequences. Bioinformatic tools can *assemble* these sequence fragments into the complete sequence of our biological sample (Section 3.3).

All sequencing methods make errors, and bioinformatic tools aim to correct them. Further

Box 5: Polymerase chain reaction (PCR)

The *polymerase chain reaction (PCR)* is a chemical reaction that increases the number of physical copies of a template DNA molecule. It uses template molecules, primers, free nucleotide molecules (dNTPs = deoxyribonucleoside triphosphates), and the *DNA polymerase enzyme*.

Primers are short, single-stranded DNA segments complementary to parts of the template molecules (usually the terminal ends). One way to ensure complementarity is by designing the primers based on selected segments from the known sequence of a genome. Another way is to attach adaptors (short double-stranded DNA fragments) to the ends of the templates.

The DNA polymerase enzyme is an enzyme that occurs in all lifeforms and plays a central role in DNA replication by synthesising the complement to the template DNA. A bacterial polymerase is used for PCR, even though it is very error-prone. Other polymerases, such as human polymerase, have much lower error rates but cannot be used since they denature at the high temperatures required for the reaction.

A PCR starts by heating the mixture to a high temperature (90°C) to split the double-stranded template molecules into single-stranded templates. The mixture is then slowly cooled down to allow the primers to bind to the templates. The polymerase enzyme then binds to where the primer ends and the template overhangs. Polymerase synthesises the complement to the template DNA by progressively adding free nucleotides, complementary to the template molecule, to the primer's end. This progresses until the end of the template, creating a new double-stranded DNA molecule. Each original double-stranded template results in two new double-stranded DNA molecules, each including one original single strand and a complementary copy.

This cycle of heating up, cooling down, and extending primers is repeated several times until the template molecule is amplified in sufficiently high numbers.

details on error profiles and appropriate correction methods can be found in Ross et al. (2013), Elloumi (2017) and Heydari et al. (2017).

Currently, we distinguish three generations of sequencing frameworks. Each is based on slightly different principles of detecting the signal from individual nucleotides. A rough summary of their performance is given in Section 2.1. In the remainder of this chapter, we discuss these frameworks in detail.

2.2 First generation: Sanger sequencing

Sanger sequencing is the oldest sequencing method that still has the lowest error rate[1] (Sanger, Nicklen and Coulson 1977). As such, it continues to serve as the gold standard of sequencing. Researchers opt for this method if an observation needs to be verified, for example, when a genome variant needs to be confirmed as real and not a sequencing error.

[1] The Nobel prize in chemistry in 1980 was awarded to Frederick Sanger and his colleagues for their contribution to the effort to decipher the genetic code.

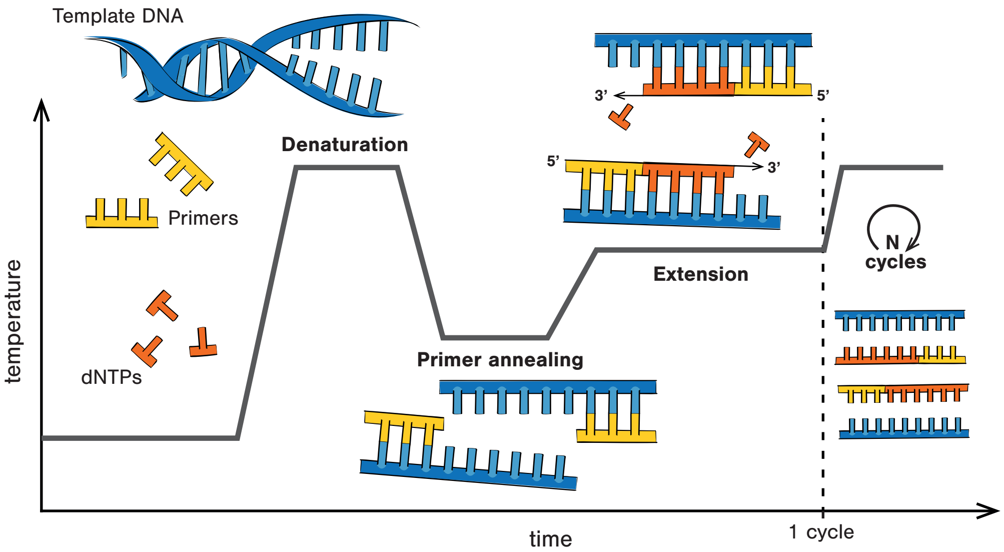


Figure 2.1: Polymerase chain reaction (PCR) consists of n cycles of replication of DNA templates. In each cycle, the mixture of template DNA, primers, and free nucleotides (the dNTPs) is first heated to a high temperature to denature the double-stranded DNA molecules and break them into single-stranded templates. The mixture is then cooled to allow primers to bind to the template (anneal). The polymerase will use the resulting overhangs to extend the complementary DNA strand and complete the template duplication. At the end of the process, 2^n copies of the template will have been produced.

The throughput of this method goes up to only 100kbp per hour (kbp = kilo base pairs, 1 000bp), and the read length is up to 1 000 nucleotides. Typically, one can simultaneously sequence up to 96 different templates per run (limited by the number of wells on the reaction plate). One significant disadvantage of Sanger sequencing is that it requires much laborious manual work, such as using bacteria for fragment separation and individual PCR for different parts of the genome.

2.2.1 Separation

Sanger sequencing typically uses bacteria for template separation, primarily *Escherichia coli*. Recombinant DNA molecules, each composed of a vector (e.g. a plasmid) plus an inserted DNA fragment (the template), are introduced into a solution of bacteria. The bacteria take up the recombinant DNA molecules with the inserted template. The concentration of recombinant DNA molecules in the solution is selected so that most bacteria pick up one molecule.

Metric	Generation		
	First	Second	Third
Speed (bp[†]/hour)	10^5	3×10^{10}	3×10^8
Typical read length (bp)	1 000	2×150	30 000[*]
Reads produced in parallel	96	10^6	10^5
Amplification step	yes	yes	no
Error rate	low	medium	medium-high

[†] Base pairs (number of nucleotides).

[*] Although the average size of fragments remains around 30 000bp, more than 2 million base pair reads have been reported for Oxford Nanopore (Amarasinghe et al. 2020).

Table 2.1: Key characteristics of the three sequencing technology generations. Note that the numbers are averages for the particular generation.

The bacteria then multiply on a plate, each creating a single colony. When a bacterium multiplies, it passes on a copy of its genetic information to each daughter cell. Thus, each daughter cell gets a copy of the chromosomal DNA and, in addition, inherits a copy of the recombinant DNA (Lodish et al. 2000). Each colony represents a clonal population stemming from a single bacterium that took up a single template. The templates should stay separated, meaning the individual colonies must be picked manually from the plate one by one and put into separate reaction tubes.

2.2.2 Amplification

Next, we need to isolate the DNA from the bacteria and amplify it further. Further amplification of the template is necessary as the sequencing method is not sensitive enough to detect the signal on a small number of template copies. For this, bacteria are lysed (their membranes are destroyed), and the proteins are denatured at high temperatures. Amplification is performed with PCR (see Box 5 on page 29), employing primers complementary to the ends of the vector into which the DNA template was inserted.

2.2.3 Sequencing

Once the amplification is finished, multiple copies of the same template are present within each tube. Reading the template sequence requires visually distinguishing nucleotides and their positions in the sequence. Again, this is based on PCR and starts from a primer. However, the PCR is carried out with a mixture of normal nucleotides (normal dNTPs) and chain-terminating inhibitors (dideoxyribonucleoside triphosphates — ddNTPs). The inhibitors are chemically modified to disallow further nucleotide attachment and act as polymerisation terminators. These molecules also contain radio-labelled phosphor to make them easier to distinguish. The first version of this method contained four steps (Sanger, Nicklen and Coulson 1977). The content of each tube is separated into four parts so that four sequencing reactions can be performed separately. In the first reaction, the normal nucleotides are mixed with the chain-terminating adenine A variant (ddATP). The chain is stopped whenever this nucleotide is built in by the polymerase. After several hours, this reaction produces a mixture of short sequences of different lengths. This reaction is repeated with chain-terminating variants of C, T, and G. The resulting mixtures are then processed using gel electrophoresis. Shorter sequences move farther on the gel plates while longer sequences move less, and their positions can be observed due to the radioactive labelling. This allows reading out the terminating nucleotide for sequences of different lengths, allowing the template sequence to be reconstructed. Figure 2.2A illustrates this process.

2.2.4 Fluorescent labels

The chain-terminating nucleotides from the original setup were later replaced by nucleotides labelled with fluorophores to speed up the process and eliminate the dangerous radio-labelling. Fluorophores are molecules that emit light at a specific wavelength when excited. As the DNA polymerase progresses along the template, it attaches different nucleotides to the end of the primer. Again, as for the radio-labelled phosphor, since the tube contains a mixture of unlabelled and labelled chain-terminating nucleotides, the labelled inhibitors will terminate the primer extension at random positions as the template is being copied, producing many partial copies of the template with varying lengths and terminal labels. In this case, we only require one rather than four reactions as we can distinguish the four labelled nucleotides.

Afterwards, the newly produced partial copies of a template are separated on a gel using an electric current. Again, the shorter copies will travel further in the gel as they are smaller and thus move more easily through the gel, while the longer ones are larger and do not travel as far. The different copies have discrete lengths, so fragments of the same length will end up close to one another, with shorter and longer fragments being lower or higher on the gel, respectively.

Next, a laser is used to excite the light-emitting molecules along the length of the gel. A detector then reads the corresponding labelling of the nucleotide at each particular position in the gel based on the colour (wavelength) of the light that is emitted. For each position, one can determine the most common colour and thus reconstruct the template sequence. Figure 2.2B shows a schematic of the process described here.

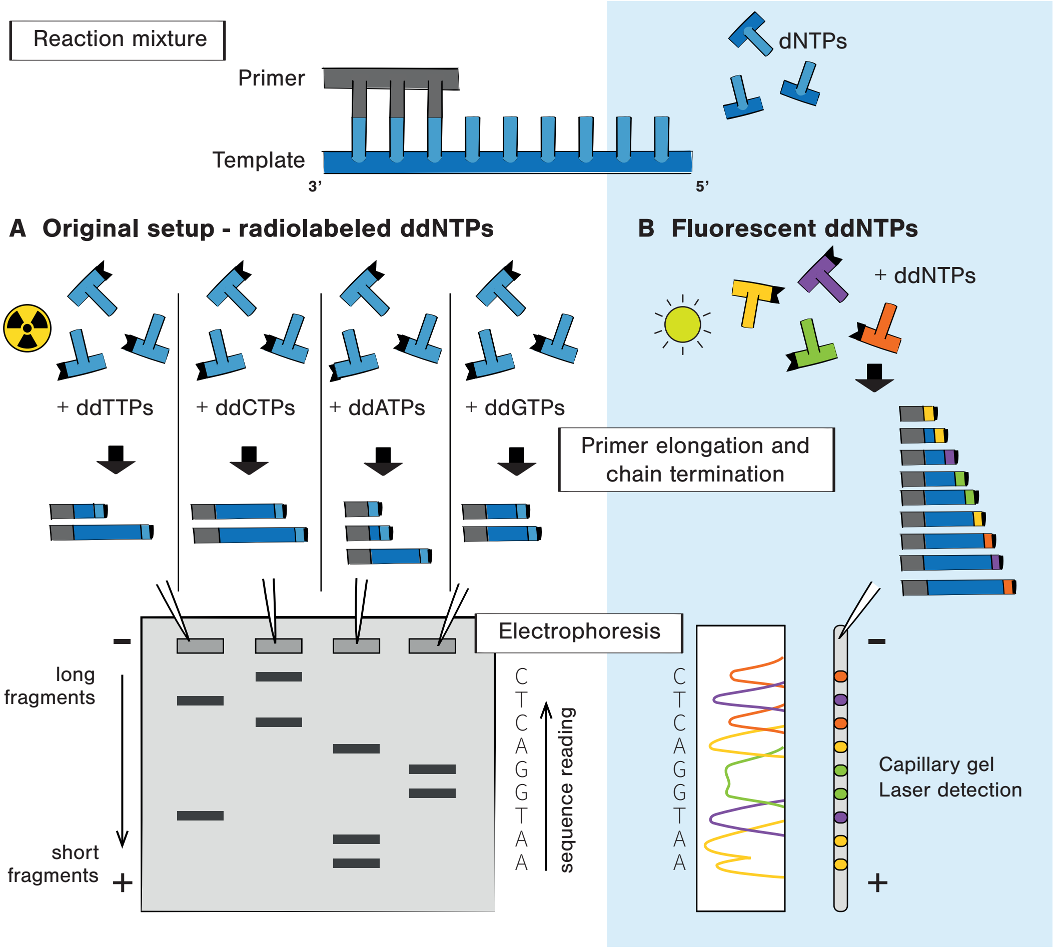


Figure 2.2: Sanger sequencing follows the principle of a PCR but with labelled chain-terminating nucleotides. **A** The original Sanger method uses four separate solutions of radio-labelled chain-terminating nucleotides. These are mixed with normal nucleotides in four parallel reactions. Each time a chain-terminating nucleotide is incorporated, DNA replication is stopped producing a sequence of a different length. When run in gel electrophoresis, shorter DNA fragments migrate farther than longer ones and the sequence can be read out from the bands. **B** Sanger sequencing with fluorescently-labelled chain-terminating nucleotides follows the same principle but uses a single mixture of unlabelled nucleotides and fluorescently-labelled inhibitors. During the PCR, partial template copies of different lengths are created, each containing a light-emitting inhibitor marking the base at the end. The templates are again separated on a gel, and the sequence is read out by exciting the terminating fluorophores with a laser.

2.3 Second generation: next-generation sequencing

Next-generation sequencing (NGS) was the name given to a group of new sequencing techniques developed in the mid to late 1990s. These methods are also called *high-throughput methods* as they can be parallelised to up to 6 000 million reads per run, with each read of up to 650 nucleotides in length. Depending on the exact platform and run mode selected, sequencing speed can go up to 30Gbp per hour (1Gbp = 10^9bp). As with Sanger sequencing, NGS methods have separation, amplification, and sequencing steps, though each step is performed in high throughput. The massively parallel nature of NGS methods makes them much cheaper[2] (have lower per nucleotide cost) than Sanger sequencing and allows multiplexing of samples. In multiplexing, DNA samples from different individuals can be mixed during a sequencing run. To distinguish which sequence came from which individual, the DNA samples are barcoded with a known short, unique stretch of DNA before mixing.

NGS employs high-throughput approaches for initial separation and amplification of the templates, and we discuss the various approaches below. By not relying on bacteria, NGS overcomes the time-consuming step of separation and amplification in Sanger sequencing. However, the NGS technologies suffer from a decrease in the accuracy of the reads as the sequence gets longer. Therefore, we can only obtain reliable sequences of a few hundred base pairs. The reads are very accurate at the beginning of the sequence, but the error rate grows as the read proceeds towards the end of the template.

Although third-generation sequencing exists, there is still a high demand for NGS sequencers as NGS is a stable and easily available technology. Furthermore, many types of post-processing methods are available to improve sequencing error correction, alignment, and analysis of the reads.

Four NGS platforms have been released to the market: SOLiD (formerly Applied Biosystems, then Thermo Fisher, now discontinued), Ion Torrent by Thermo Fisher, 454 by Roche (discontinued in 2016), and HiSeq/MiSeq/…by Illumina, the most widespread NGS technology at the time of writing.

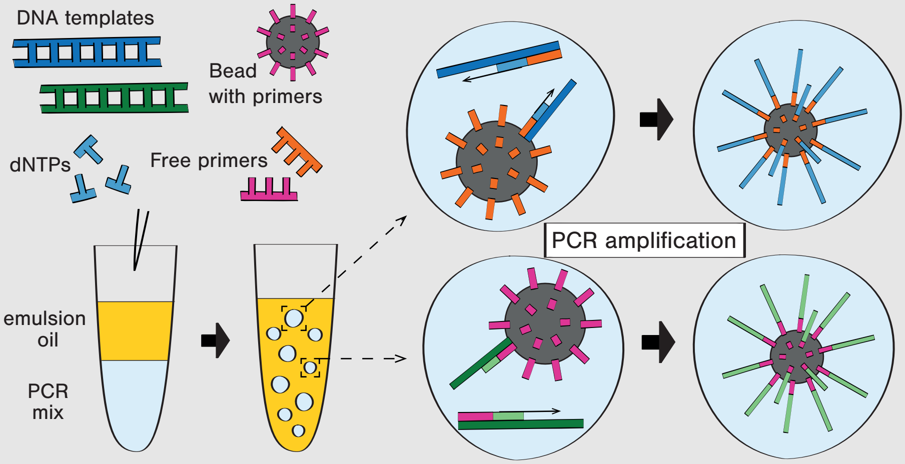
2.3.1 Pyrosequencing with 454 from Roche (discontinued in 2016)

The 454 uses emulsion PCR (see Box 6 on page 35) to separate and amplify the template of interest. After emulsion PCR, each droplet should contain a single bead covered in the amplified template. The beads are then distributed on a plate with wells, accommodating exactly one bead per well. The beads are fixed and separated by falling into the wells on the plate. This way, the amplified templates are physically separated, which is the step for which Sanger sequencing requires bacterial cloning and manual colony picking.

[2] NGS sequencing is cheaper if you already own the sequencing machine. There is a tradeoff between the single-run cost and the investment to buy the instrument.

Box 6: Emulsion polymerase chain reaction

Emulsion PCR is a PCR where each template is amplified in its own chamber. Emulsion PCR uses a water phase that contains the DNA templates, free nucleotides, DNA polymerase enzyme, and beads (tiny sphere molecules) with attached primers. The water phase is mixed with oil, creating droplets that act as PCR reaction chambers. Then a PCR is performed as explained in Box 5 on page 29, with the single-stranded templates attaching to a primer on the bead or to a free primer so that the complement of the template is synthesised. The mixture is set up so that most droplets contain a single bead with a single template. Since the individual chambers do not interact during the PCR, at the end of the amplification reaction, each bead will be covered in many copies of the template that paired up with it in the reaction chamber. However, sometimes, a single droplet may contain one bead but more than one template, causing the bead to emit mixed signals during sequencing. Since only a few beads in the whole reaction contain such mixed signals, this drawback of the method is greatly overpowered by the amount of useful information produced.



The sequencing is performed as the sequence is synthesised (sequencing by synthesis). The wells are sequentially flooded with unmodified A, C, G or T nucleotides (each flooding is performed with only one of the four nucleotides), which are washed away before the next nucleotide is introduced. Light is produced and recorded when that particular nucleotide is incorporated. In particular, through a series of reactions, the nucleotide's double phosphate group (pyrophosphate) activates another molecule called luciferin, which emits light. This procedure, called *pyrosequencing*, is shown schematically in Figure 2.3.

This method was the first next-generation sequencing technique to be released to the market. However, it may encounter problems displaying high error rates when dealing with sequences containing nucleotide homopolymers (single nucleotide repeats), introducing multiple identical nucleotides. Signals generated from a high repeat number are difficult to distinguish from similar but slightly different repeat lengths (such as 8- and 7-nucleotide repeats).

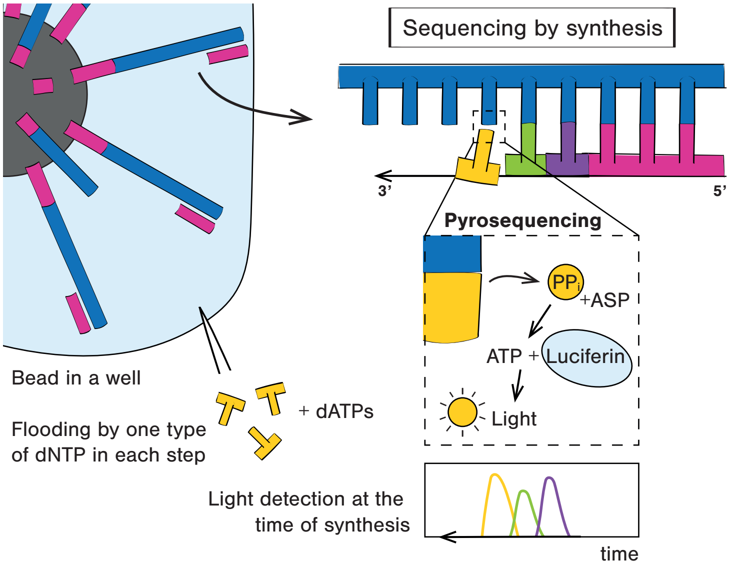


Figure 2.3: Pyrosequencing. The double phosphate group is detached as each new nucleotide is attached to the template sequence. A light-emitting molecule, luciferin, is activated through a series of reactions, and the light is recorded to indicate which nucleotide was incorporated.

Furthermore, the 454 sequencer itself and its consumables were quite expensive. However, this method produced the longest reads available among next-generation sequencing methods.

2.3.2 SOLiD from Thermo Fisher (discontinued)

SOLiD sequencing also uses emulsion PCR (see Box 6 on page 35) for separation and amplification of the template of interest. The beads are then transferred to a chemically treated slide where they bind to the surface.

The sequencing itself is performed by ligation, running in the direction opposite to DNA synthesis. First, the primer binds to the template. The reaction then uses an enzyme ligase to join the primer to 8 nucleotide-long stretches of DNA (octamers). The binding octamers are complementary to the template. These stretches are fluorescently labelled based on the first two nucleotides. The last nucleotide of the stretch is modified so that no nucleotide can attach further. Schematically, this stretch can be encoded as XYNNNZZZ, where X and Y are the nucleotides determining the fluorescent label, N are the degenerate bases that are used for indentation and Z are the universal bases where the last one carries the fluorescent label and disallows the attachment of further octamers. After reading the light emitted by the fluorescent markers labelling the combination of X and Y, one step of the process is completed. The process continues by cutting off the last three bases of the octamer (ZZZ) to allow the next marked octamer to attach. Once the whole template is processed in this fashion, we have traversed its sequence in steps of 5 bases at a time: two bases are read, and three bases form the yet unknown gaps.

After one such run, the newly created sequence is erased, and a new primer is attached. The new primer attaches to a position in a template that is shifted forward by a single nucleotide with respect to the previous primer attachment position. The primer extension, sequence erasing, and primer shifting are repeated four more times. Each base of the whole template sequence is interrogated (read) twice due to the two-base colour-coded sequencing and the primer shifting. The complete sequence in question can be reconstructed from the combined light signals from each run. Figure 2.4 schematically shows the procedure of one primer extension run.

A significant advantage of this method is that each of the bases is interrogated twice during a single run, thereby decreasing the number of potential errors. The main drawback is a low coverage in AT-rich repetitive regions, and regions where a stretch rich in A or T nucleotides (e.g. TAT, TAAAA, or TGTT) is repeated several times (Harismendy et al. 2009).

2.3.3 Ion Torrent from Thermo Fisher

Same as 454 and SOLiD, Ion Torrent uses emulsion PCR (see Box 6 on page 35) to amplify and separate the templates. Each of the beads is placed in a well on a plate, and the sequencing is performed as the sequence is synthesised. In contrast to 454, the recorded signal when a nucleotide is added is not that of the pyrophosphate but rather a hydrogen ion (H^+). Each time a new nucleotide is attached to the synthesised sequence, a hydrogen ion (H^+) is released, which is sensed by the semi-conductor plate located below each of the wells. The H^+ ion is only released (and thus recorded) when the correct nucleotide binds in one of the four successive floods of different nucleotides. No light-emitting nucleotides and no optical measurements are required. Figure 2.5 shows the cross-section of the well with the bead and the Ion Torrent sequencing principle.

This method has the same issue as 454, namely distinguishing homopolymer repeats. The method's main advantages are that it is quite cheap since it does not require any fluorescent molecules, lasers, or detectors, and it is also fast since there is no need for steps such as the excitation of fluorescent molecules to determine the sequence.

2.3.4 Illumina sequencers

In Illumina sequencing, the templates are separated and amplified using the solid support of a primer-covered slide rather than beads. The templates are washed over the slide and attached to the primers, ideally with enough space between the attached templates. Isothermal bridge

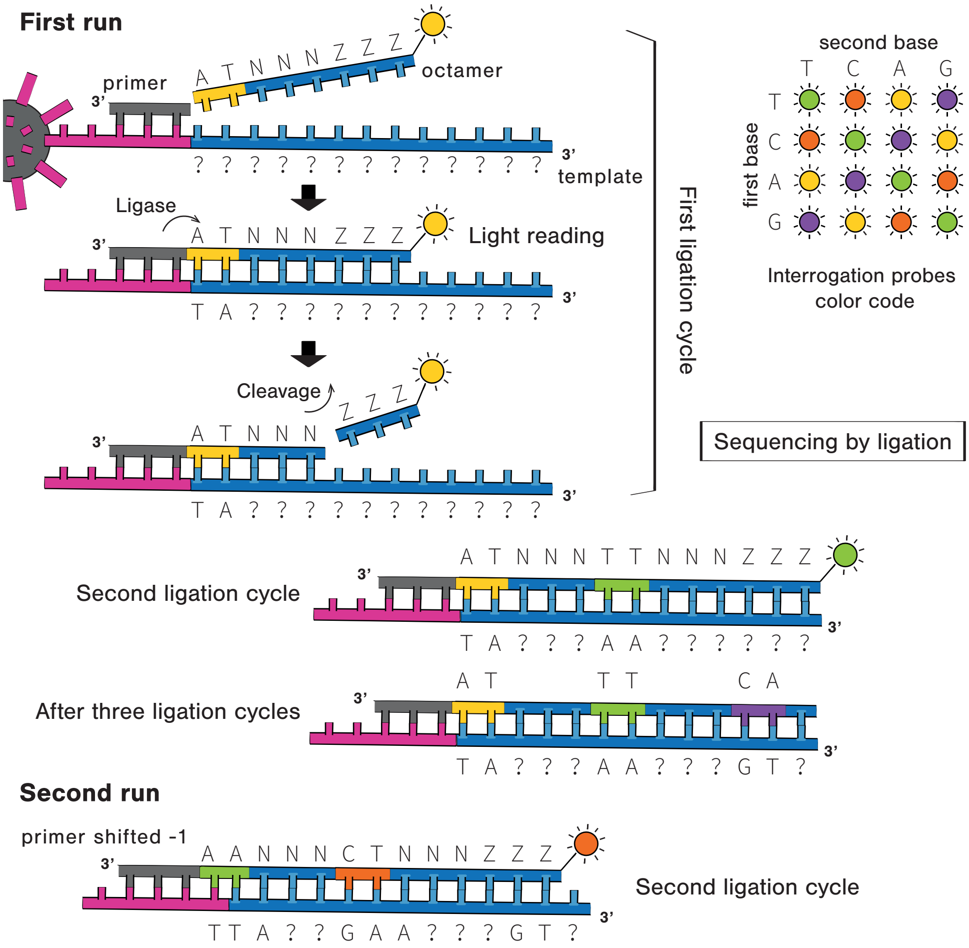


Figure 2.4: SOLiD sequencing. This figure shows the procedure for a single run-through of the sequencing in SOLiD. A nucleotide stretch of length 8 attaches to the primer, which allows the light signal from the first two marked nucleotides to be detected. Then, the last three nucleotides are cut off, and another 8-nucleotide stretch attaches. This way, two bases get read, and three get skipped in a single run. After a run is complete, the primer is shifted by one nucleotide, and the procedure is repeated. Overall, five runs are performed, ensuring each base is read twice.

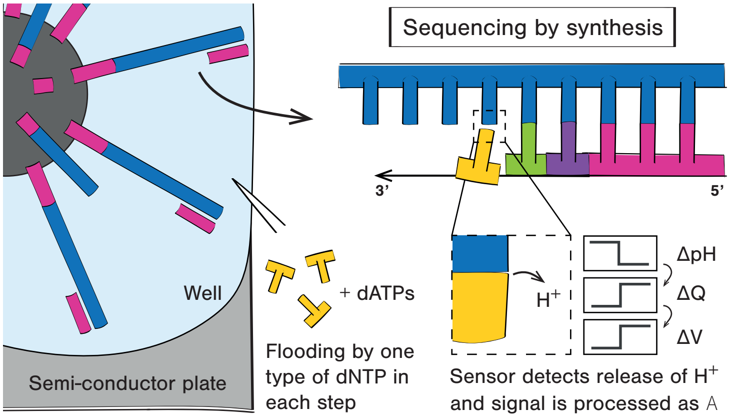

Sequencing by synthesis
3'
5'
Well
+ dATPs
Semi-conductor plate
Flooding by one type of dNTP in each step
H+
ΔpH
ΔQ
ΔV
Sensor detects release of H+ and signal is processed as A

Figure 2.5: Ion Torrent sequencing. The wells containing template-covered beads are sequentially flooded with different nucleotides. Each time a new nucleotide is attached to the template, a H^+ ion is released. The ion is detected by the sensor plate located beneath the bottom of the well containing the template-covered bead.

amplification is then performed to create clusters of copies of the same template around the initial attached template. Figure 2.6 shows this process.

Sequencing is performed by synthesis using nucleotides labelled with different fluorophores, modified to act as temporary inhibitors of synthesis[3]. As the nucleotides are distinctly labelled, we can wash a mixture of all nucleotides over the slide and then record the emitted light at each spot (cluster) of the slide.

Illumina offers a wide array of machines with differing levels of throughput. The company currently offers the best price per base pair and the fastest sequencing time. As with many other methods using lasers and optical devices for signal detection, the accuracy of reads decreases towards the end of the template.

2.4 Third generation: Single-molecule sequencing

The third generation of sequencing methods is the latest wave of sequencing methodologies, allowing for much longer read lengths. At each step of the NGS sequencing process, all copies of the template on a bead or in a cluster on a slide receive a new nucleotide; thus, the signal comes from not one but many copies of the template simultaneously and is amplified this way.

[3] The HighSeq platforms use four different fluorophores, the NextSeq and MiniSeq platforms only use two. Two nucleotides are labelled with one of these fluorophores each. The third nucleotide is labelled with both, and the fourth nucleotide is not labelled at all. Thus, upon incorporation, the fourth nucleotide emits no light.

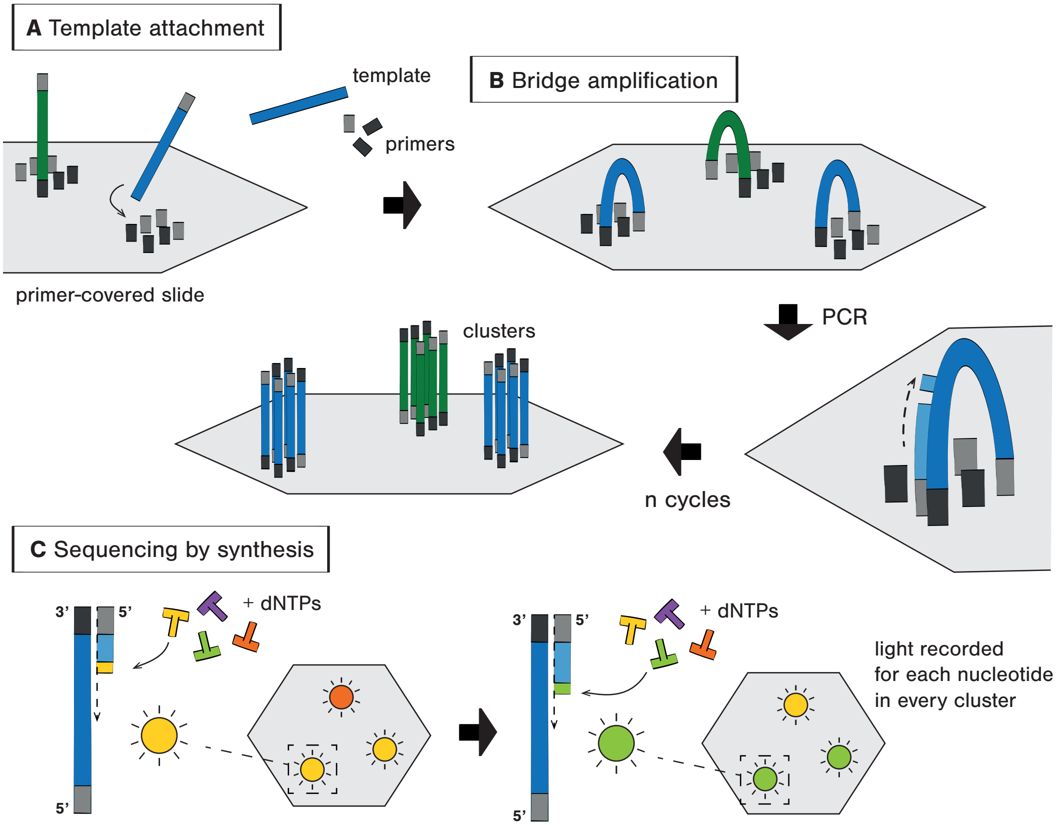


Figure 2.6: Illumina sequencing. **A** Illumina sequencers use a solid support instead of beads. **B** A primer-covered slide is washed with templates, which attach and then form clusters as they are amplified. This process achieves what the bacterial colonies were doing in Sanger: separating different templates. **C** During sequencing, nucleotides modified to act as temporary inhibitors of synthesis and labelled with different fluorophores are incorporated.

In contrast, third-generation sequencers do not need signal amplification. Instead, these aptly named *single-molecule sequencing methods* can detect the signal of a single molecule. This means they do not require target template amplification, which is one of the main sources of errors for all previously mentioned methods. It is the main advantage of third-generation sequencing and the main difference compared to the second-generation. For example, a single viral genome could be sequenced without amplification, given that its genome is shorter than the sequencer's read length. This genome would constitute a single template to be sequenced. Note that if we want to sequence a genome that is longer than the sequencer's read length, library preparation still requires an amplification step to include templates spanning the whole genome into the library. After sequencing the templates, assembly tools can help reconstruct the whole genome sequence based on the individual template reads.

Third-generation sequencers employ various sequencing principles. Here, we will briefly sketch strand synthesis (used in PacBio RS/PacBio Sequel by Pacific Biosciences) and pore-based sequencing (used in MinION/GridION/PromethION by Oxford Nanopore) since these have been the most widely used at the time of writing. These methods allow for average read lengths of about 30kbp and offer high throughput with parallelisation of up to 10 million reads per run. The disadvantage of the third-generation sequencers is that they are still quite expensive due to the high-resolution detection systems needed to achieve single-molecule resolution, and the reads generally contain a higher number of sequencing errors compared to NGS methods. However, both the costs and error rates of these methods are continuously decreasing.

2.4.1 PacBio RS from Pacific Biosciences

PacBio uses single-molecule real-time (SMRT) sequencing technology, which does not require PCR. A plate with wells is used to separate templates, with a polymerase attached at the bottom of each well. The nucleotides used for synthesis are labelled with different fluorescent markers. As the complement of the template is synthesised, the well with the polymerase is illuminated from the bottom. The fluorophore attached to the nucleotide being incorporated into the growing DNA strand emits light, which is immediately read by the detector. PacBio uses proprietary technologies such as zero-mode waveguide cells that help guide and focus the light on the bottom of the individual wells to allow for precise detection of the incorporation of a single nucleotide. SMRT is a real-time technology since DNA synthesis and nucleotide detection occur simultaneously, allowing the sequencer to record nucleotide incorporation into the sequence in real-time during synthesis (Figure 2.7). In contrast, all second-generation technologies added a nucleotide but then had to wait for the sequencer to detect the signal before the next addition, meaning synthesis was interrupted for measurement.

This sequencing method produces raw reads with many errors, as the polymerase has a relatively high error rate ($10 - 15\%$). However, this can be remedied by circularising the template and letting the synthesis continue for several rounds (a process called circular consensus sequencing, CCS). Since the polymerase makes random and not position-dependent errors, the errors will occur at different positions in each read. Bioinformatic software can thus distinguish errors from true nucleotides by the majority rule and exclude them from the final result. The trade-off is that CCS reads are typically shorter than the more error-prone single-pass long reads.

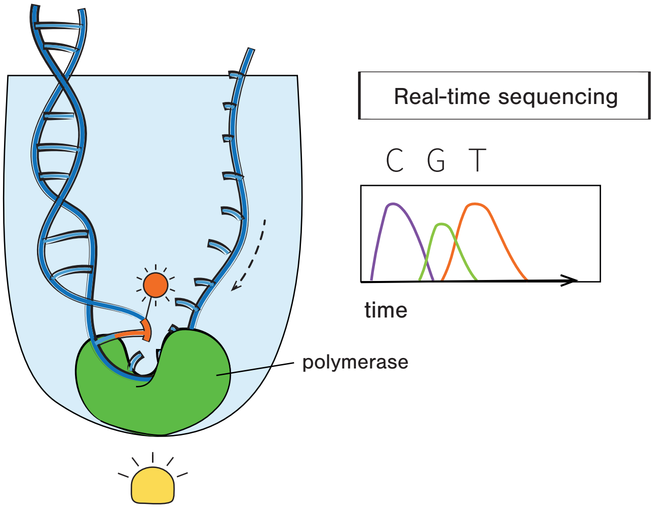


Figure 2.7: PacBio sequencing. The reaction well with the polymerase enzyme fixed to its bottom is illuminated from underneath. A labelled nucleotide, complementary to the template strand, enters the reaction site of the polymerase enzyme. The nucleotide is immobilised for a tiny fraction of time, sufficient for the dye attached to the nucleotide to be illuminated and to emit a nucleotide-specific signal. The dye is cleaved off the nucleotide upon DNA synthesis, and the nucleotide is attached to the growing DNA strand.

2.4.2 MinION/GridION/PromethION from Oxford Nanopore

The MinION/GridION/PromethION sequencers from Oxford Nanopore all use nanopore technology for sequencing. The nanopores can be biological, such as pore proteins within a lipid membrane, or engineered, such as graphene. Single-stranded DNA is fed through the pore by an enzyme attached to the edge of the pore (shown in Figure 2.8). An electric current passes through the material in which the pores are embedded, and as a nucleotide passes through the pore, the electric current changes. The drop in current differs depending on the size and chemical composition of the molecule passing through the pore. As such, the sequencer can tell which particular stretch of nucleotides passes through the pore at a specific time. The method can even distinguish between methylated and non-methylated versions of a particular nucleotide as it changes the time it takes to pass through the pore.

Like PacBio, the nanopore technology detects the sequence in real-time as the nucleotides pass through the pore. Combined with special bioinformatic tools, this can allow highly efficient species or resistance classification (within 10 minutes) (Břinda et al. 2020).

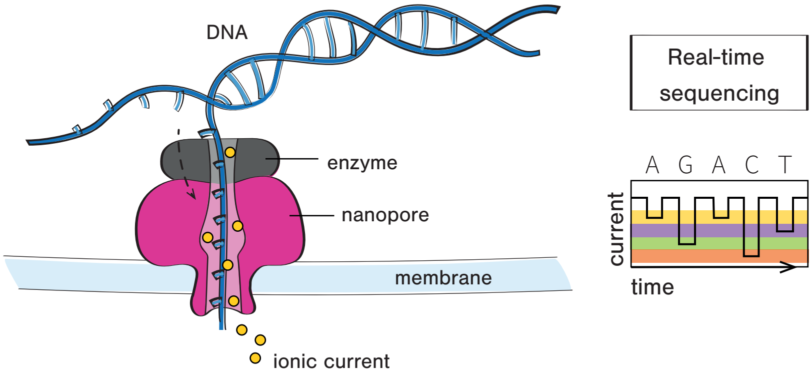


Figure 2.8: Nanopore sequencing. The enzyme attached to the nanopore feeds the intact template through the pore. The identity of the nucleotides is then read by the change in the electric current caused by the template passing through the nanopore.

3 Sequence alignments

In the previous chapter, we discussed how to obtain a genetic sequence from a biological sample of an individual (e.g. a species or an infected host). This individual is typically a representative of a larger population. When the biological unit of interest is a species, typically, each species is represented by a single sequenced individual from that species. For pathogens within a host, the consensus (in a sense, the average) sequence of the pathogen population is often used to represent that infected host.

After sequencing DNA from different individuals, we want to compare these sequences to find similarities and differences. However, before doing this, we need to identify the regions in the sequences that correspond to the same position in the genome for all the individual sequences. The result of such a process is called an *alignment*. More precisely, alignments consist of several sequences from different individuals (or some biological unit as listed in Section 1.1.1) at the nucleotide or amino acid level. Sequences are called *aligned* if each character (nucleotide or amino acid) within a sequence has an assigned unique position. In an alignment, sequences are typically displayed in different rows, and characters with the same assigned position are displayed in the same column. Each column in the alignment is referred to as a *site* in the alignment. If a particular sequence has no character at a particular site, we view that specific position as a gap (represented by a hyphen −).

In this context, we introduce three important terms, *orthologues*, *paralogues*, and *homologues*. Two nucleotides in different sequences are orthologues (or are orthologous to one another) if they have a shared ancestral nucleotide and were separated through speciation or the corresponding birth event when considering biological units different from species. Similarly, two nucleotides in different sequences are paralogues if they have a shared ancestral nucleotide but are separated through gene duplication. The set of both paralogues and orthologues is referred to as homologues. The homologous nucleotides may differ in the two sequences due to errors in replication, but that does not change their homologous nature.

When constructing an alignment, we typically aim to ensure that each site consists of orthologous nucleotides, and thus, characters are ideally assigned to sites in the following way. Suppose we know the phylogeny describing the ancestry of the different sequenced individuals; then, the characters across sequences are assigned to the same site if they correspond to the same ancestral character in the most recent common ancestor sequence. The characters at a particular site may differ across individuals due to point mutations. An insertion (or a repeat or gene duplication) in an individual adds a site (or multiple sites) to the alignment. All sequences that do not have this new fragment will have gaps at those sites. Similarly, the individuals with a deletion will have gaps representing the deleted characters.

In the case of a recombination, each character is traced back to the root sequence via its ancestor prior to the recombination. In the case of an inversion of a part of the sequence, the order of characters in the sequences changes when they are aligned. For example, consider characters $k, \ldots, k+m$ in sequence A, and an inversion of characters $k, \ldots, k+m$ in sequence B. In the alignment, assume the characters $k, \ldots, k+m$ of sequence A are assigned to sites $k, \ldots, k+m$. Then, the characters $k, \ldots, k+m$ of the inverted sequence are assigned to sites $k+m, \ldots, k$; that is, their order is reversed in the alignment.

However, since we typically do not know the true phylogeny nor the history of the evolution of the sequences on that phylogeny, we aim to find an alignment that is hopefully close to the true unknown alignment using a certain model or some optimisation criterion, typically not involving a phylogeny[1].

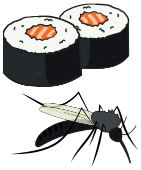
There are two types of alignments: *pairwise alignments* where only two sequences are aligned, and *multiple sequence alignments (MSA)*, where a set of multiple sequences are aligned to each other. This chapter will discuss exact methods for pairwise alignment and heuristics for pairwise alignment and MSA. Exact means that the method provides the best output under some optimality criterion (but that optimality criterion may not necessarily ensure that the output alignment has correctly assigned the orthologues). A heuristic will not necessarily provide the optimal answer under the optimality criterion; however, it is typically much faster than the exact method. Heuristics for pairwise alignments are widely used for matching one sequence against a big library, for example, using the BLAST algorithm (*Basic Local Alignment Search Tool*). Heuristics for MSA are the methods of choice for aligning sequences from multiple individuals since exact methods are computationally intractable. Besides the dot-matrix method (described in Section 3.1.1), presented alignment tools share the property that sequences are aligned by maintaining the order of characters in each sequence and simply adding gaps to certain positions in each sequence. In particular, these methods cannot account for inversions. While this may bias downstream results, it is still the standard way of obtaining alignments.

As discussed in the previous chapter, the reads obtained from sequencing are often shorter than the genomic region we want to compare across individuals. Thus, we end this chapter by describing how the presented alignment algorithms can be used to assemble the reads to obtain the genetic sequence of the biological sample. Once the complete genetic sequences are assembled, we can proceed with the regular alignment and analysis techniques.

In summary, this chapter outlines the process of going from genetic samples from a population of individuals to the alignment as illustrated in Figure 1.2, and describes how to find genetic sequences similar to a particular sequenced one using the data mining approach BLAST.

We finish this introduction with an empirical example of a pairwise alignment.

[1] There are methods capable of jointly inferring both the phylogeny (the evolutionary history) and the alignment from a set of unaligned sequences. While this approach is statistically superior to the approaches presented in this section, it is very computationally demanding even for small datasets (Redelings and Suchard 2005; Suchard and Redelings 2006; Redelings and Suchard 2007; Redelings 2014).



```
NGTTDQVDKIVKILNEGQIASTDVVEVVVSPPYVFLPVVKSQLRPEIQVAAQNCW
||....!..!.|!|..|.!.:.  .||||. | .!|.:.!|||...! ||||||!
NGDKASIADLCKVLTTGPLNAD__TEVVVGCPAPYLTLARSQLPDSVCVAAQNCY
```

Figure 3.1: Pairwise alignment of partial amino acid sequences of the triose-phosphate isomerase enzyme — responsible for efficient energy production in cells — in mosquito and rice.

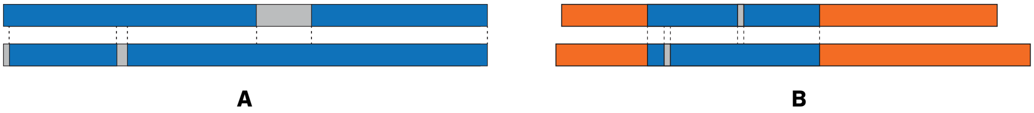


Figure 3.2: Schematic of **A** global and **B** local alignment: aligned parts of the sequences are in blue, non-aligned parts of the sequences in orange and gaps in grey.

Example: Triose-phosphate isomerase The enzyme triose-phosphate isomerase is essential for efficient energy production in cells. Since it fulfils a vital role, it can be found in most eukaryotes. Figure 3.1 shows an alignment of the amino acid sequence of this protein in two very different species: rice and mosquito.

Since this is an amino acid sequence, the alignment does not only take into account perfect matches (represented by the symbol |) but also positions where the two corresponding amino acids have similar chemical properties and can play similar roles in the protein (represented by the symbols !, : and . for strong, medium and low similarity, respectively). Only 36.4% of the positions in the two sequences match perfectly; however, most positions contain similar amino acids, meaning this is a so-called "well-conserved" protein.

3.1 Pairwise alignments

A pairwise alignment can be a *local alignment* or a *global alignment*, as shown in Figure 3.2. A global alignment contains both sequences aligned from start to end, whereas a local alignment only aligns sub-sequences.

Like all alignments, pairwise alignments can be between different types of sequences: protein-protein alignments as in the triose-phosphate isomerase example, DNA-DNA, RNA-RNA, and even DNA or RNA with protein. In the latter, there is no one-to-one correspondence between characters in the DNA/RNA sequence and characters in the protein sequence: one amino acid in the protein sequence corresponds to a codon (3 nucleotides) in the DNA or RNA sequence (see Section 1.2.7). This makes dealing with insertions and deletions more complicated. Another difficulty is that multiple codons can encode for the same amino acid, a phenomenon called *codon degeneracy*. We focus here on DNA-DNA alignments.

There are several strategies to build pairwise alignments, which we will cover in the following sections:

(i) the *dot-matrix method*, which strictly speaking does not lead to an alignment but visualises similarity (Gibbs and McIntyre 1970) (Section 3.1.1),

(ii) the *exhaustive method* of listing all possible alignments and scoring them according to some scoring scheme (Section 3.1.2), and then returning the one or multiple alignments with the highest score (Section 3.1.3),

(iii) the *Needleman-Wunsch algorithm* for global alignment (Needleman and Wunsch 1970), and its equivalent, the *Smith-Waterman algorithm* for local alignment (Smith and Waterman 1981), which rely on *dynamic programming* to obtain the alignments with the highest score (Section 3.1.4),

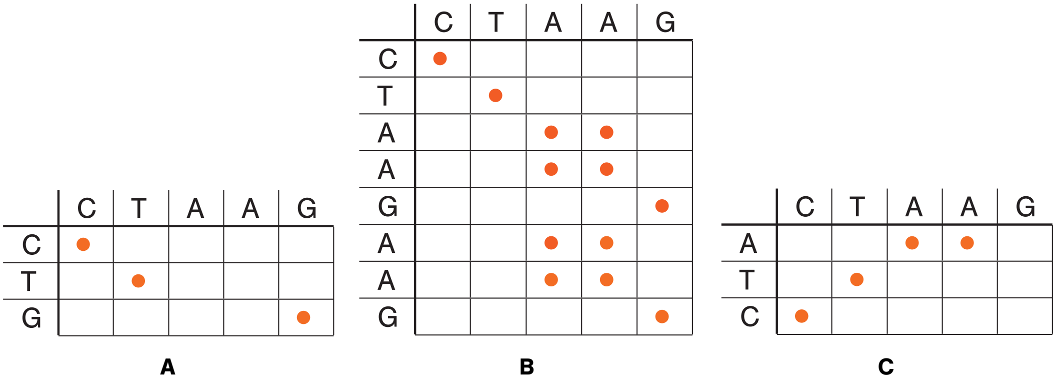
(iv) the BLAST algorithm as a heuristic for local alignments (Section 3.1.5).

Note that the exhaustive method and the Needleman-Wunsch algorithm both output the alignment with the highest score. The latter is superior in speed due to dynamic programming. Nevertheless, we also present the exhaustive method to illustrate the basic ideas behind score-based alignments.

3.1.1 Dot-matrix method

The dot-matrix method helps visualise the similarity between two sequences (Gibbs and McIntyre 1970). In this method, the two sequences are arranged in a matrix such that the rows represent the characters of one sequence and the columns represent the characters of the second sequence. Each position in the matrix in which the character in the column matches the character in the row is marked with a dot in the respective field. All other positions are left blank.

As shown in Figure 3.3, this method makes it easy to identify important features visually: *indels (insertions and deletions)* are represented as gaps in the matrix (Figure 3.3 **A**). Repeats are visible as repeated pattern blocks that are shifted horizontally or vertically (for example, in Figure 3.3 **B**, there is a 3 × 3 block that is shifted vertically). Inversions are visible as reflected diagonal patterns (Figure 3.3 **C**). However, this method does not return an alignment; rather, it visually highlights areas of sequence similarity.



	C	T	A	A	G
C	●				
T		●			
G					●

A

	C	T	A	A	G
C	●				
T		●			
A			●	●	
A			●	●	
G					●
A			●	●	
A			●	●	
G					●

B

	C	T	A	A	G
A			●	●	
T		●			
C	●				

C

Figure 3.3: Dot-matrices illustrating important features when comparing two sequences: **A** insertions and deletions (indels), **B** repeats and **C** an inverted sequence.

3.1.2 Scoring schemes

Alignment methods rely on a scoring scheme or a stochastic model to evaluate alignments and pick the one with the highest score or likelihood. The methods presented here require a scoring scheme. In Chapter 5, we will introduce stochastic models for sequence evolution, which can also be used for alignment methods using statistical approaches such as *maximum likelihood* (the concept of maximum likelihood is explained in Box 25 on page 116).

The choice of a scoring scheme will strongly affect the result, as any optimal alignment is only optimal under the specific scheme used to evaluate it. Most scoring schemes treat all positions in the alignment as independent: the score of the entire alignment is simply the sum of the scores at each position. In the simplest scheme, a position gets one of three possible scores depending on whether there is a match, a mismatch, or a gap at this position. Matches will increase the alignment score, whereas mismatches and gaps will decrease it. Gaps, representing insertions or deletions, are biologically less likely to happen, so they generally incur a higher penalty than mismatches. In particular, the order of characters in each sequence is maintained in the alignment, and gaps are added between characters of a sequence. In our examples, we will use the following scoring scheme:

(i) match score $= 3$;

(ii) mismatch score $= -1$;

(iii) gap score $= -2$.

More complex scoring schemes can be used; for example, one extension makes the gap scores dependent on the length of the gap. The reasoning here is that opening a gap is an unlikely event, but the longer an existing gap is, the easier it is to extend it. Another possible extension is to use substitution matrices (see Chapter 5), which imply different scores for mismatches depending on which two characters are aligned at a position.

3.1.3 Exhaustive method

The exhaustive method lists all possible alignments for two sequences, scores them, and chooses the ones with the highest score. Note that a site with a gap in both sequences will not appear in any alignment, as such sites are uninformative and will only decrease the score. This method will always return the alignment with the highest score, but the computation is very slow as it depends on the total number of possible alignments. We will prove in Theorem 3.1.1 that the number of possible alignments grows very fast (in fact, exponentially) with sequence length.

Theorem 3.1.1. *For two sequences of lengths m and n, $m \geq n$, the number of possible alignments is $\sum_{k=0}^{n} \binom{m+k}{k}\binom{m}{m+k-n}$.*

Proof. Let $A = a_1\, a_2\, \ldots\, a_m$ and $B = b_1\, b_2\, \ldots\, b_n$ be the two given sequences of lengths m and n respectively, $m \geq n$.

Assume there are $k \leq n$ gaps introduced in the sequence A in the alignment ($k > n$ would lead to at least one site at which both sequences would have a gap, which is not allowed). The alignment is then of length $m + k$, which means there are $\binom{m+k}{k}$ possible gap placement locations in sequence A.

Similarly, there are k' gaps introduced in sequence B in the alignment, $m + k = n + k'$, $k' = m+k-n$. There cannot be a gap in both sequences at the same position in the alignment. Thus, the k' gaps in sequence B need to be aligned with characters of sequence A, which gives $\binom{m}{k'} = \binom{m}{m+k-n}$ possibilities for placing those gaps.

Finally, we need to sum over all possible values of $k \leq n$ to account for all possible numbers of gaps in the alignment, arriving at $\sum_{k=0}^{n} \binom{m+k}{k}\binom{m}{m+k-n}$. □

This number grows very fast with the length of both sequences: for relatively short sequences of lengths $m = n = 100$, there are already 2.05×10^{75} possible alignments.

Thus, the exhaustive method is not practical for anything other than extremely short sequences.

3.1.4 Dynamic algorithms

The Needleman-Wunsch (Needleman and Wunsch 1970) and Smith-Waterman (Smith and Waterman 1981) algorithms are algorithms for global and local alignments, respectively. The Needleman-Wunsch algorithm returns the same alignments as the exhaustive method: the alignments with the highest score. The Smith-Waterman algorithm computes the best local alignments, which returns the highest-scored sub-sequence alignments. This means that the resulting local alignments cannot start or end with gaps, as that would decrease the score. The local and global alignments of two sequences may be identical but usually differ. Both algorithms score alignments faster (polynomial runtime) and thus more efficiently than the exhaustive method (exponential runtime).

To speed up the alignment process, we observe that many possible alignments share the same start or share some other regions. For example, the alignments $\begin{array}{l}\texttt{ATTG-C}\\\texttt{A-TACC}\end{array}$ and $\begin{array}{l}\texttt{ATT-GC}\\\texttt{A-TACC}\end{array}$ are identical in the first three positions $\begin{array}{l}\texttt{ATT}\\\texttt{A-T}\end{array}$. This means that computing the score for several alignments from scratch would involve calculating the score of the same sub-alignment. Therefore, we can speed up the score calculations by storing the scores of the sub-alignments instead of recomputing them every time.

Moreover, we can go one step further and directly determine the highest-scoring sub-alignments (pairwise alignments of sub-sequences), save their scores, and use them to obtain the highest-scoring complete alignment. This technique of using a solution to a subproblem to solve the complete problem is called *dynamic programming*. This general technique is often used in bioinformatics, phylogenetics, and phylodynamics, as seen throughout this book.

The Needleman-Wunsch and Smith-Waterman algorithms apply the concept of dynamic programming by storing the scores of best sub-alignments in a matrix that spans the two sequences. This matrix contains information in the form of arrows to reconstruct the highest-scoring complete alignment based on the values in the matrix. In principle, the two algorithms are very similar and differ only in minor details. We will explain the general ideas by focusing on the Needleman-Wunsch algorithm and then highlighting the differences in the Smith-Waterman algorithm.

3.1.4.1 Needleman-Wunsch algorithm

Suppose we want to align two sequences, A $= a_1\, a_2 \ldots a_m$ and B $= b_1\, b_2 \ldots b_n$, where a_i denotes the character at position i in sequence A and b_j the character at position j in sequence B. To determine the highest-scoring alignment, we write down a matrix, H, of dimensions $(m+1) \times (n+1)$. Recall that the cell (i, j) in a matrix is the entry in the ith row and jth column. For convenience, we count the rows and columns starting from 0. Rows 1 to m represent the characters of sequence A, and columns 1 to n column represent the characters of B, as shown in Figure 3.4 **A**.

In the Needleman-Wunsch algorithm, the entry in the scoring matrix H_{NW} at position (i, j), denoted by $H_{\mathrm{NW}}(i, j)$, represents the score of the highest-scoring alignment of sequences $a_1\, a_2\, \ldots\, a_i$ and $b_1\, b_2\, \ldots\, b_j$. Note that $i = 0$ corresponds to an empty sub-sequence of A, and $j = 0$ corresponds to an empty sub-sequence of B. The initial condition is $H_{\mathrm{NW}}(0, 0) = 0$; that is, the score of an empty alignment is 0.

Now we iteratively fill out the matrix for all i, j, where $H_{\mathrm{NW}}(m, n)$ will contain the score of the highest-scoring alignment of sequences A and B. Assume we have calculated $H_{\mathrm{NW}}(k, l)$ for all $k \leq i$ and $l \leq j$ where $k + l < i + j$. Next, we want to calculate $H_{\mathrm{NW}}(i, j)$. There are three choices for the possible alignment of sub-sequences $a_1\, a_2\, \ldots\, a_i$ and $b_1\, b_2\, \ldots\, b_j$, but we really only need to consider the last characters of the sub-sequences, a_i and b_j:

(i) if a_i and b_j are aligned and the best score of the alignment of the sequences $a_1\, a_2\, \ldots\, a_{i-1}$ and $b_1\, b_2\, \ldots\, b_{j-1}$ is $H_{\mathrm{NW}}(i-1, j-1)$, then we need to add the score for a (mis-)match (depending on whether $a_i = b_j$ or not), $s(i, j)$, to $H_{\mathrm{NW}}(i-1, j-1)$ to get $H_{\mathrm{NW}}(i, j)$;

(ii) if b_j is aligned with a gap in sequence A and the best score of the alignments of sequences $a_1\, a_2\, \ldots\, a_i$ and $b_1\, b_2\, \ldots\, b_{j-1}$ is $H_{\mathrm{NW}}(i, j-1)$, then we need to add the gap penalty, w, to $H_{\mathrm{NW}}(i, j-1)$ to get $H_{\mathrm{NW}}(i, j)$;

(iii) if a_i is aligned with a gap in sequence B and the best score of the alignments of sequences $a_1\, a_2\, \ldots\, a_{i-1}$ and $b_1\, b_2\, \ldots\, b_j$ is $H_{\mathrm{NW}}(i-1, j)$, then we need to add the gap penalty, w, to $H_{\mathrm{NW}}(i-1, j)$ to get $H_{\mathrm{NW}}(i, j)$.

As we are looking for the highest score $H_{\mathrm{NW}}(i, j)$ at position (i, j), we need to calculate all three possibilities and choose the sub-alignment leading to the highest value.

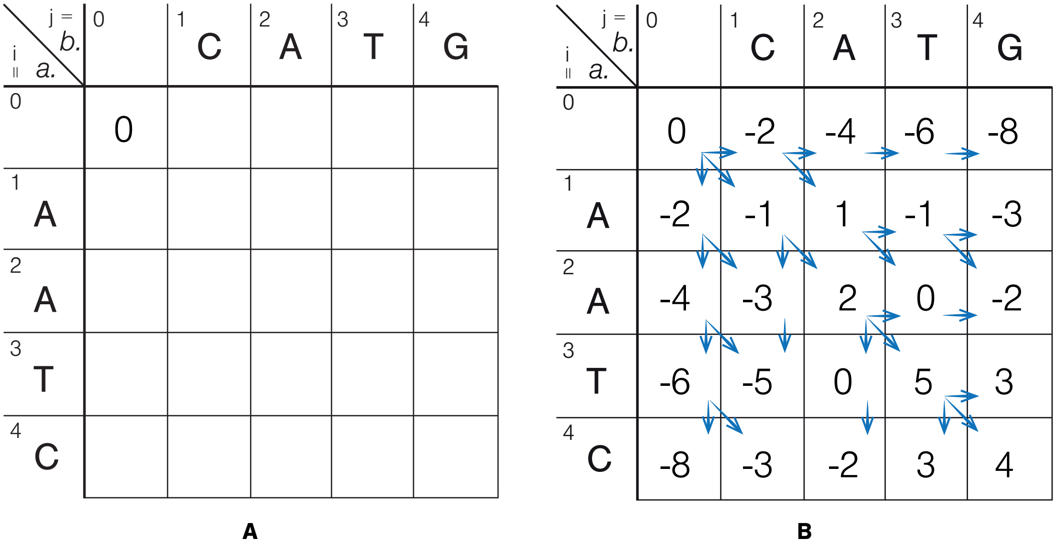
In mathematical terms, we can express these rules as

$$H_{\mathrm{NW}}(i, j) = max \begin{cases} H_{\mathrm{NW}}(i-1, j-1) + s(i, j) & \text{(mis-)match (case 1)}, \\ H_{\mathrm{NW}}(i, j-1) + w & \text{gap in sequence A (case 2)}, \\ H_{\mathrm{NW}}(i-1, j) + w & \text{gap in sequence B (case 3)}, \end{cases} \tag{3.1}$$

where $s(i, j)$ is the score of a match if $a_i = b_j$ and the score of a mismatch otherwise, and w is the score of a gap.

For each value $H_{\mathrm{NW}}(i, j)$, we additionally note whether the best score comes from cases 1, 2, or 3 in Equation (3.1). If the best score at position (i, j) resulted from case 1, the addition of a (mis-)matched character pair $a_i,\ b_j$ to the sequence represented at position $(i-1, j-1)$, a diagonal arrow from field $(i-1, j-1)$ to field (i, j) is added. If the best score was achieved by adding a gap to A (case 2), we draw an arrow to the right, from field $(i, j-1)$ to (i, j). If the best score was achieved by adding a gap to B (case 3), we draw a down arrow from field $(i-1, j)$ to (i, j). Several options can give the same score, in which case all corresponding arrows are noted. An example is shown in Figure 3.4 **B**.

Once the score matrix has been filled, the best global alignment score is found in the bottom right corner of the matrix, in position (m, n). The alignment can be reconstructed by following



i = a. \ j = b.	0	1 C	2 A	3 T	4 G
0	0				
1 A					
2 A					
3 T					
4 C					

A

i = a. \ j = b.	0	1 C	2 A	3 T	4 G
0	0	-2	-4	-6	-8
1 A	-2	-1	1	-1	-3
2 A	-4	-3	2	0	-2
3 T	-6	-5	0	5	3
4 C	-8	-3	-2	3	4

B

Figure 3.4: Example of filling out the score matrix in the Needleman-Wunsch algorithm. **A** Initial and **B** complete score matrices for the sequences $\mathrm{A} = \mathsf{AATC}$ and $\mathrm{B} = \mathsf{CATG}$ using the Needleman-Wunsch algorithm. We attribute a score of 3 to matches ($s(i,j) = 3$ if $a_i = b_j$), -1 for mismatches ($s(i,j) = -1$ if $a_i \neq b_j$) and penalise a gap with $w = -2$.

the arrows backwards to the top left corner of the matrix, position $(0,0)$, and adding the corresponding characters to the alignment as we traverse the matrix in reverse:

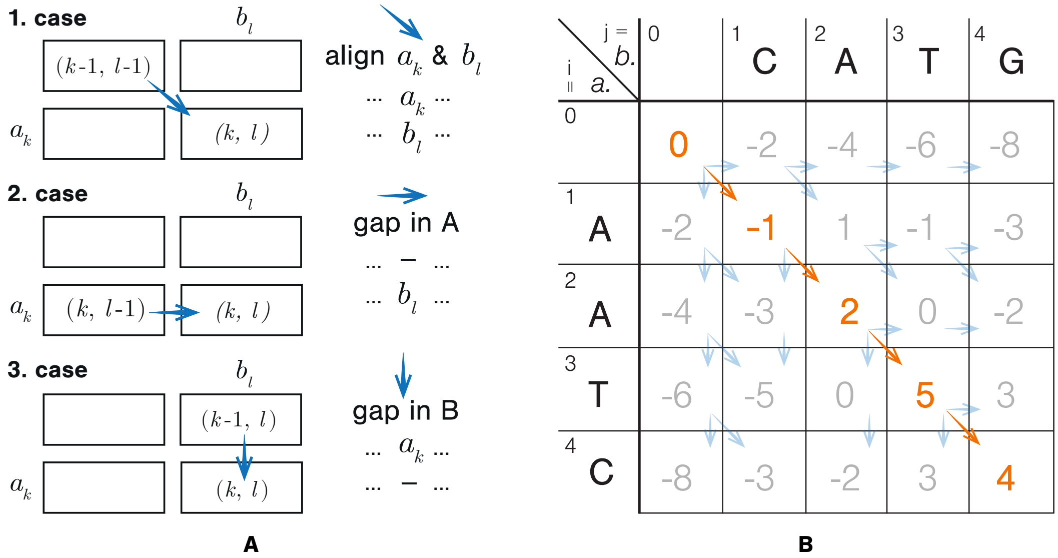
(i) if the arrow pointing to position (k,l) is diagonal, then characters a_k and b_l are added to the alignment, aligned with each other;

(ii) if the arrow pointing to position (k,l) is a right arrow, then the character b_l is added to the alignment, aligned with a gap in sequence A;

(iii) if the arrow pointing to position (k,l) is a down arrow, then the character a_k is added to the alignment, aligned with a gap in sequence B.

Figure 3.5 **A** illustrates how to read the arrows and build the alignment in reverse order. Figure 3.5 **B** demonstrates how to go backwards in the example from Figure 3.4 and obtain the global alignment. Note that if more than one arrow points towards position (k,l), each way leads to an alignment with the same (highest) score.

The number of calculations required to obtain the best global alignment can be obtained by noting that we need the following steps: (i) to fill out each cell of the matrix, we evaluate three products (perform three calculations) over which we take the maximum (see rule for $H_{\mathrm{NW}}(i,j)$); (ii) we fill out $(m+1) \times (n+1)$ matrix entries; (iii) we reconstruct the alignment by



i = a. \ j = b.	0	1 C	2 A	3 T	4 G
0	0	-2	-4	-6	-8
1 A	-2	-1	1	-1	-3
2 A	-4	-3	2	0	-2
3 T	-6	-5	0	5	3
4 C	-8	-3	-2	3	4

Figure 3.5: Constructing the alignment using the Needleman-Wunsch algorithm. **A** Schematic of how to read the arrows in the score matrix. **B** Score matrix of the example in Figure 3.4 with the way through the matrix. The final global alignment is $\begin{smallmatrix}\texttt{CATG}\\ \texttt{AATC}\end{smallmatrix}$ and has a score of 4.

tracing back at most $m+n$ arrows. This amounts to $3\times(n+1)\times(m+1)+(n+m)$ calculations. Thus, for the Needleman-Wunsch algorithm, we have the polynomial runtime $\mathcal{O}(nm)$ (see Box 7 on page 55). Note that for $n = m = 100$, this is on the order of 10^4 steps, while the exhaustive method with an exponential runtime is on the order of 10^{75} steps, illustrating the power of the dynamic programming method to make complex problems computationally tractable.

3.1.4.2 Smith-Waterman algorithm

The Smith-Waterman algorithm (Smith and Waterman 1981) returns the best local alignment of two sequences, only aligning sub-sequences. More precisely, it finds the best alignment for sub-sequences $a_i,\dots,a_k$ and $b_j,\dots,b_l$, where i,j,k,l are chosen such that the score of the alignment is the maximal score. The algorithm follows a scheme similar to the Needleman-Wunsch algorithm. However, it differs in the initialisation of the score matrix, the calculation of the scores, and the start/end of the reverse alignment reconstruction from the score matrix. We will now describe how to fill out the scoring matrix, H_{SW}. As before, we denote the two sequences with A and B and their lengths with m and n, respectively.

Box 7: Landau symbol and algorithmic runtimes

The *Landau symbol* or big $\mathcal{O}$ notation is used to provide an approximation of the asymptotic behaviour of a function. The notation $\mathcal{O}$ was first used by the German mathematician Paul Bachmann and widely spread by the German mathematician Edmund Landau. Thus, this notation is also referred to as Bachmann-Landau notation and goes back to the German expression "Ordnung von", which translates to "order of".

A function $f(x)$, where x is a real value, has the order of $g(x)$, a positive real-valued function if there exist some constants C and x_0 such that $|f(x)| \leq C \times g(x)$ for all $x > x_0$. We denote this property by $f(x) = \mathcal{O}(g(x))$.

Throughout this book, we use the $\mathcal{O}$-notation to report asymptotic runtimes. Runtime on the order $\mathcal{O}(g(N))$ means that there exist some constants C and N_0 such that the number of calculations required to solve the problem, f, is $f \leq C \times g(N)$ for all $N > N_0$, where N is the input data size.

We say that an algorithm has *polynomial runtime* if $g(N) = N^k$, where k is some constant that is independent of N. We say that an algorithm has *exponential runtime* if $g(N) = e^{N^k}$. Since polynomials grow much slower in N than exponentials, algorithms with polynomial runtime are much faster than those with exponential runtime. *Linear runtime* is a special case of polynomial runtime where $k = 1$.

Note that $f(x) = O(g(x))$ establishes an upper bound; for example, an algorithm with runtime $\mathcal{O}(N)$ is always also an algorithm of runtime $\mathcal{O}(e^N)$.

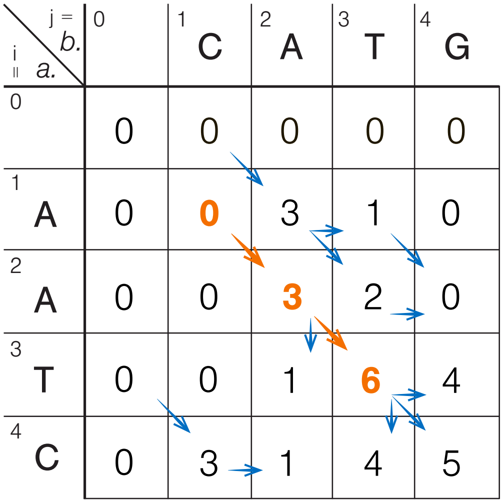
Initialisation The score matrix H_{SW} has the dimensions $(m+1) \times (n+1)$, exactly as H_{NW}. However, the entries of row 0 and column 0 are set to 0.

Score matrix The score matrix is then successively filled based on the scoring function:

$$H_{\mathrm{SW}}(i,j) = max \begin{cases} 0 & \text{(stop)}, \\ H_{\mathrm{SW}}(i-1,j-1) + s(i,j) & \text{(mis-)match (case 1)}, \\ H_{\mathrm{SW}}(i,j-1) + w & \text{gap in sequence A (case 2)}, \\ H_{\mathrm{SW}}(i-1,j) + w & \text{gap in sequence B (case 3)}, \end{cases} \tag{3.2}$$

where $s(i,j)$ is the score of a match if $a_i = b_j$ and the score of a mismatch otherwise, and w is the score of a gap.

Building the final alignment The alignment reconstruction starts at the position with the highest score, (k,l) (rather than the bottom right field (m,n)). This highest score is the score of the local alignment, which ends with the aligned nucleotides a_k, b_l. The alignment reconstruction proceeds similarly to the Needleman-Wunsch algorithm but stops when a position $(i-1, j-1)$ with a score of 0 is reached. The alignment thus starts with the aligned nucleotides a_i, b_j. Figure 3.6 shows the Smith-Waterman score matrix for the example sequences AATC



j = i =	0	1 C	2 A	3 T	4 G
0	0	0	0	0	0
1 A	0	0	3	1	0
2 A	0	0	3	2	0
3 T	0	0	1	6	4
4 C	0	3	1	4	5

Figure 3.6: Example of the score matrix using the Smith-Waterman algorithm for sequences $A = \texttt{AATC}$ and $B = \texttt{CATG}$. We again use the scoring scheme $s(i, j) = 3$ if $a_i = b_j$, $s(i, j) = -1$ if $a_i \neq b_j$, and $w = -2$. The best local alignment is $\begin{matrix}\texttt{AT}\\ \texttt{AT}\end{matrix}$.

and CATG (the same sequences we used in the Needleman-Wunsch example). The best local alignment is $\begin{matrix}\texttt{AT}\\ \texttt{AT}\end{matrix}$ whereas the best global alignment was $\begin{matrix}\texttt{CATG}\\ \texttt{AATC}\end{matrix}$.

This procedure ensures that we find the alignment of sub-sequences with the highest score; that is, we find the best local alignment. Like in the Needleman-Wunsch algorithm, several local alignments can have the same highest score and thus are equally good. Again, the runtime is $\mathcal{O}(nm)$.

3.1.5 Heuristic alignments: BLAST

Imagine we obtained a genetic sequence and want to discover whether this or a similar sequence has been found before. This problem requires a large number of pairwise sequence alignments and can arise in different contexts:

(i) imagine a patient who suffers from symptoms that cannot be unanimously assigned to a specific disease; the only clue we have is a pathogen sequence extracted from the patient, and we want to find out which pathogen it is;

(ii) imagine we found a gene in an organism, but we are not sure which function this gene encodes;

(iii) or, imagine that we have sequenced the genome from a particular individual but do not know which species it belongs to.

These cases have in common that one obtains an unknown sequence, in this context, also referred to as a *query sequence*. We want to compare the query sequence to already known sequences from a huge database — referred to as *library*. In particular, we want to find homologues of the query sequence. Known characteristics of the homologues, such as the corresponding species or function, allow us to hypothesise about the characteristics of the query sequence and, consequently, about the individual from which the sequence was obtained. Differences between the query sequence and its homologues inform us about genotypic variation between the underlying individuals. Investigating the differences can be done either using GWAS-like approaches (Chapter 4) or by reconstructing evolutionary history in the form of the phylogeny (Chapter 6).

The comparison of the obtained sequence to the sequences in the library essentially means that we calculate local pairwise alignments between our obtained sequence and each library sequence. Although the dynamic programming algorithms introduced in Section 3.1.4 are much more efficient than the exhaustive approach in calculating these pairwise alignments, they are still too slow for scanning a big library of sequences.

One solution is to use *Basic Local Alignment Search Tool* BLAST (`https://blast.ncbi.nlm.nih.gov/Blast.cgi`), a heuristic word algorithm first published by Altschul et al. (1990). BLAST takes advantage of the fact that two similar sequences contain local alignments of short sub-sequences with high alignment scores. Two completely different sequences do not contain such local alignments and can be excluded early in the search process.

Let the query sequence be of length n. In general, the algorithm has three steps:

1. compiling a list of words of a certain length based on the query sequence,
2. scanning the database (the library) for hits (matches) to these words, and
3. extending found matches.

Step 1: Words of short length are determined based on the query sequence of length n. A word is a sequence of length k (also called a k-*mer*). A simple option to generate such a list is to break down the query sequence into k-mers by starting at the first position and moving one character for each new word. The word list then contains $n - k + 1$ words.

For example, let us consider the following query sequence of length $n = 10$:

```
NYEFILKWCL
```

This sequence can be cut into the following 3-mers:

```
NYE YEF EFI FIL ILK LKW KWC WCL
```

Thus, only words that occur in the original sequence are considered.

When searching for a nucleotide sequence, the standard setting for k-mer length is $k = 28$; a match is scored with $+1$ and a mismatch with -2.

Another option to generate the word list — which is suggested in the original paper by Altschul et al. (1990) for amino acid sequence searches — is to consider all k-mers that align with some part of the query sequence and have an alignment score that is bigger than a pre-defined value T. The authors used the PAM-120 matrix[2] to determine the score of the k-mers (Altschul et al. 1990). Nowadays, the default substitution matrix for the most common BLAST implementation for amino acid sequence alignment is BLOSUM62 (BLOcks SUbstitution Matrix, shown in Figure 3.7)[3]. For the following explanation of the BLAST algorithm, we will also use the BLOSUM62 matrix for amino acid substitutions.

Let us again consider the sequence `NYEFILKWCL` as an example, and let us assume that we only consider words of length $k = 3$ that score at least $T = 18$ when aligned to the query sequence. The score is calculated using the BLOSUM62 matrix (Figure 3.7). For example, let us compute the score $S_{\mathtt{NYE}}$ of the 3-mer `NYE` when aligning to `NYEFILKWCL`:

```
NYE-------
NYEFILKWCL
```

This local alignment has the score $S_{\mathtt{NYE}} = 6 + 7 + 5 = 18$. If we only include words with a score of at least $T = 18$, `NYE` would be added to the list.

However, the best alignment of the substring `EFI` looks like this:

```
--EFI-----
NYEFILKWCL
```

It scores $S_{\mathtt{EFI}} = 5 + 6 + 4 = 15$ and would not be added to the list of words, even though it is an exact substring of the query sequence.

The word `TWC` however aligns as follows:

```
------TWC-
NYEFILKWCL
```

[2] The PAM (point accepted mutation) matrix describes substitution scores from one amino acid to another. The PAM-1 matrix lists the substitution scores in case 1% of the sequences were different, PAM-250 lists these scores in case 20% amino acids were different. PAM matrices are not symmetric, so, for example, a substitution from `G` to `F` can have a different score than `F` to `G` (Dayhoff, Schwartz and Orcutt 1978).

[3] The BLOSUM family contains symmetric amino acid substitution matrices. The alignment score of two sequences that are 62% identical sum up to 1. These matrices were introduced in Henikoff and Henikoff (1992).

C	9																			
S	−1	4																		
T	−1	1	5																	
P	−3	−1	−1	7																
A	0	1	0	−1	4															
G	−3	0	−2	−2	0	6														
N	−3	1	0	−2	−2	0	6													
D	−3	0	−1	−1	−2	−1	1	6												
E	−4	0	−1	−1	−1	−2	0	2	5											
Q	−3	0	−1	−1	−1	−2	0	0	2	5										
H	−3	−1	−2	−2	−2	−2	1	−1	0	0	8									
R	−3	−1	−1	−2	−1	−2	0	−2	0	1	0	5								
K	−3	0	−1	−1	−1	−2	0	−1	1	1	−1	2	5							
M	−1	−1	−1	−2	−1	−3	−2	−3	−2	0	−2	−1	−1	5						
I	−1	−2	−1	−3	−1	−4	−3	−3	−3	−3	−3	−3	−3	1	4					
L	−1	−2	−1	−3	−1	−4	−3	−4	−3	−2	−3	−2	−2	2	2	4				
V	−1	−2	0	−2	0	−3	−3	−3	−2	−2	−3	−3	−2	1	3	1	4			
F	−2	−2	−2	−4	−2	−3	−3	−3	−3	−3	−1	−3	−3	0	0	0	−1	6		
Y	−2	−2	−2	−3	−2	−3	−2	−3	−2	−1	2	−2	−2	−1	−1	−1	−1	3	7	
W	−2	−3	−2	−4	−3	−2	−4	−4	−3	−2	−2	−3	−3	−1	−3	−2	−3	1	2	11
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W

Figure 3.7: BLOSUM62 matrix as derived in Henikoff and Henikoff (1992).

It scores $S_{\mathsf{TWC}} = -1 + 11 + 9 = 19$ and would be added to the list of words even though it is not an exact substring of the query sequence. This way of generating the list of words based on similarity score is used in protein BLAST with the standard settings $k = 6$ and $T = 10$. The authors of the original publication state that "[i]f a little care is taken in programming, the list of words can be generated in time essentially proportional to the length of the list." (Altschul et al. 1990).

Query sequence: NYEFILKWCL

Sequence from database: NYEFGGTWCL

Step	Word	Score
Initial word	TWC	$5+11+9=25$
Expansion 1	TWCL	$5+11+9+4=29$
Expansion 2	LTWCL	$-4+5+11+9+4=25$
Expansion 3	ILTWCL	$-4-4+5+11+9+4=21$

Figure 3.8: Illustration of the word extending step of BLAST. We start with the word TWC from the list (Step 1), which aligns to the query sequence with a score of 19. This word aligns to the database sequence with a score of 25 (Step 2). We extend this word with characters from the query sequence (Step 3). If the score for the extended word drops below a certain threshold from its highest score at any point, the algorithm stops. In this example, the algorithm stops if the score drops below 20% of the highest score ($20\% \times 29 = 5.8$, so below $29 - 5.8 = 23.2$), which happens after the third extension step.

Step 2: For each entry in the library, it is checked whether a local alignment with one word of the list exceeds a pre-defined score T', which can be different from the acceptance score T in the first step. Only the sequences that were aligned with a score greater or equal to T' are further processed in step 3.

Step 3: The word is successively extended to both sides by the characters from the query sequence, and with each addition of a new character, the score between this extended word and the word from the library is recalculated. The process terminates when the score drops farther than a certain value below the best score for a shorter word. See Figure 3.8 for an illustration of this process. The scores and the local alignments are reported as the output of the BLAST search.

Using the standard BLAST software, one can search for both nucleotide and protein sequences. The acceptance scores and the (mis-)match scores can be adapted according to specific needs. In addition, gaps can be considered. Moreover, the acceptance criterion in Step 3 is normally more involved than what is presented in Figure 3.8 (a more detailed explanation can be found in Altschul et al. (1997)). BLAST's runtime depends on the library structure and can not be precisely estimated. However, in practice, the algorithm is extremely fast and scans huge libraries within seconds. The original BLAST algorithm sped up local alignment searches by an order of magnitude (Altschul et al. 1990). Note that while BLAST has substantial speed advantages over a search where one would use the Smith-Waterman algorithm for each pair of query and library sequences, BLAST does not necessarily return the best local alignment between these two sequences based on the chosen scoring scheme (Pertsemlidis and Fondon 2001).

3.2 Multiple sequence alignments

When more than two sequences need to be aligned, we speak of a *multiple sequence alignment (MSA)*. This is a common situation since MSAs form the basis of many larger comparison studies and are central to phylogenetic tree reconstruction. A multiple sequence alignment method aims to align a set of biological sequences (DNA, RNA, proteins), taking into account evolutionary events such as mutations, insertions, and deletions. Similar to the case of pairwise sequence alignments (see Section 3.1), computing an MSA requires optimising an objective function (an MSA score) over all possible alignments. As for pairwise alignments, this typically means maximising a sequence similarity measure (the alignment score).

MSA algorithms tend to be computationally expensive. Several possible strategies for computing an MSA exist:

(i) align all sequences against a reference sequence using pairwise alignment algorithms, which requires that a reference sequence is known and that all sequences in the alignment belong to the same or very closely related species;

(ii) extend the Smith-Waterman algorithm for multiple sequences, which will produce an exact solution but requires m^k calculation steps for k sequences of length m (meaning it has exponential computational complexity in k);

(iii) use heuristic algorithms that are faster than the exact methods but do not guarantee the optimal alignment with respect to the chosen scoring system, so they require the user to check and possibly adjust the alignment.

If we lack a good reference sequence and the dataset is too large for an exact algorithm, we are left with heuristic (approximate) algorithms for multiple sequence alignment. In the next section, we introduce such a heuristic.

3.2.1 Heuristic multiple sequence alignment methods

Generalising the pairwise alignment methods by reconstructing an alignment that maximises the sum of similarities for all pairs of sequences (the sum-of-pairs or SP score) (Edgar and Batzoglou 2006) leads to an MSA estimate. However, this problem formulation lacks a rigorous theoretical foundation for why evolution would lead to the maximal sum-of-pairs score and ignores any available phylogenetic information. Importantly, finding the alignment with the best SP score is *NP-complete* (see Box 28 on page 154) and the computational time and the required memory scale exponentially with the number of sequences (Wang and Jiang 1994).

Progressive alignment algorithms are the most common heuristic to speed up MSA computation. Rather than computing all-against-all alignments, a progressive scheme aligns a set of k sequences by performing $k - 1$ pairwise alignments with, for example, Needleman-Wunsch, where the inclusion order of sequence pairs is defined by a pre-computed guide tree (Hogeweg and Hesper 1984). The guide tree is usually computed using the neighbour-joining or UPGMA algorithms (Section 6.3.1.1) on a distance matrix derived from pairwise alignments of the input sequences (Hogeweg and Hesper 1984; Higgins and Sharp 1988). The primary example of a progressive alignment method is ClustalW (https://www.genome.jp/tools-bin/clustalw) (Thompson, Higgins and Gibson 1994). Introduced in 1994, it is still widely used, although it is less scalable and accurate than more modern approaches. Further, progressive alignment methods, in general, rely heavily on the correctness of the guide tree, which can substantially impact downstream phylogenetic analysis (Tan et al. 2015).

The progressive alignment approach has been embedded in iterative strategies (*iterative alignment algorithm*), where the guide trees and alignments are re-estimated until both converge, greatly improving the accuracy. Popular methods are MUSCLE (http://www.ebi.ac.uk/Tools/msa/muscle/) (Edgar 2004), MAFFT (https://mafft.cbrc.jp/alignment/software/) (Katoh et al. 2002), or ClustalOmega (https://www.ebi.ac.uk/Tools/msa/clustalo/) (Sievers et al. 2011).

To estimate large MSAs, the guide tree estimation has been replaced by a fast pre-clustering step using a small set of seed sequences to determine inclusion order. Examples are PartTree in MAFFT (https://mafft.cbrc.jp/alignment/software/), or ClustalOmega (https://www.ebi.ac.uk/Tools/msa/clustalo/). These methods scale well, but at the cost of reduced accuracy (Chatzou et al. 2016).

Another class of fast progressive alignment methods cluster sequences by a tree-based decomposition — for example, SATé-II finds the longest branch in the current guide tree and uses it to break the tree into two subsets. It continues to break up the longest branches in the subtrees until each subset contains at most 200 sequences. After decomposition, external tools are used to align the subsets, merge the resulting sub-alignments, and reconstruct a guide tree for the next iteration. While such tools do iterate over MSA and tree reconstruction, they rely on external reconstruction tools that do not share common evolutionary models, making their statistical properties hard to evaluate. Examples are SATé (https://phylo.bio.ku.edu/software/sate/sate.html) (Liu et al. 2009) and PASTA (https://github.com/smirarab/pasta) (Mirarab et al. 2015).

The main problem of progressive alignment methods is that they can get stuck in a local optimum of the MSA score, especially when dealing with divergent sequences (Chatzou et al. 2016). *Consistency-based alignment methods* attempt to avoid such local optima. Rather than optimising the MSA score, consistency-based algorithms maximise "agreement" (consistency) within the set of all pairwise alignments computed for the input sequences. Here, agreement means that alignments should have the same evolutionary implications, meaning that if A is similar to B and B is similar to C, then A should be similar to C. Effectively, this involves computing pairwise alignments for all input sequences and then assigning them an agreement score that signifies their compatibility with the rest of the pairwise alignments. Consistency-based methods perform progressive or iterative alignment using this agreement score rather than the MSA score. Consistency-based aligners vary in how they compute the pairwise alignments and their scores but typically yield accurate MSAs, although at a significant computational cost. Examples are T-Coffee (https://tcoffee.crg.eu/) and ProbCons (http://probcons.stanford.edu/) (Notredame, Higgins and Heringa 2000; Do et al. 2005).

A further push has been to include indel modelling into MSA estimation. One of the first such *evolution-based aligners* was PRANK (http://wasabiapp.org/software/prank/) (Löytynoja and Goldman 2005). PRANK still uses a progressive approach to aligning sequences along a guide tree. However, it models insertions and deletions in an evolutionarily meaningful way that maintains character homology[4] and uses the phylogenetic information coming from the guide tree, resulting in better alignments overall (Löytynoja 2014).

Lastly, the latest promising development is co-estimating alignments and phylogenetic trees while modelling indel evolution. The difficulty in such approaches is that they require a tractable evolutionary model that would allow estimating both objects. Such methods are computationally highly intense and so far are applicable only to small datasets; however, the amount of data such methods can handle has been consistently growing in the last few years. As of now, there are two Bayesian approaches available — BAli-Phy (https://www.bali-phy.org/), which is in active development and can handle up to 100 sequences (Redelings 2021), and StatAlign (https://statalign.github.io/), which can process up to 30 sequences (Novák et al. 2008). Another project in active development is a frequentist joint alignment and tree co-estimation tool (Pečerska, Gil and Anisimova (2021), JATI (https://github.com/acg-team/JATI)) based on a tractable model of sequence evolution with single-character indels (Bouchard-Côté and Jordan 2013).

In general, it is worth making multiple alignments with different software and checking the consistency between them. This avoids proceeding with downstream analyses with an MSA that is only a local optimum rather than a globally good estimate.

This section discussed evolutionary-based alignment approaches aiming to align homologous nucleotides. We end by noting that a large number of structural alignment approaches are available. They aim to describe structural similarity, for example, in proteins, based on their tertiary structure. These kinds of alignments can be used on sequences separated by large evolutionary distances; however, they may display similarities resulting from convergent evolution rather than shared evolutionary history. Since this book focuses on evolutionary history, we do not discuss these approaches further.

[4] Homologous characters necessarily have the same ancestor, meaning that each position in an alignment can only be described by a single insertion event.

3.3 From sequencing reads to genome sequences

So far, we have studied how to align sequences from different individuals. However, an individual's genome is typically too large to be sequenced all in one piece. Instead, sequencing technologies produce reads, which are incomplete snapshots of the underlying sequence of interest (see also Chapter 2). There are two main avenues to reconstruct the original sequence from the sequencing reads (assembly): if there is a good reference genome available, one can align the reads against this reference using a short-read alignment method; otherwise, *de novo* assembly is needed. Performant tools all rely on particular data structures for efficient assembly; we refer the reader to the provided references for details on these data structures.

3.3.1 Short-read alignment methods

Many short-read alignment and assembly methods exist, and the choice of method typically depends on the sequencing technology used, as well as the required trade-off between speed and sensitivity. Second- and third-generation sequencing techniques produce such a great amount of data that typically, heuristic algorithms need to be used, specifically methods that are optimised for speed and memory usage. The methods make assumptions about the expected length, number, and error profile of the reads, which are specific to the sequencing technology used.

Generally, short-read alignment methods align short-reads against the reference genome. The methods typically use a multistep procedure: the first step uses a heuristic technique to find a small subset of possible mapping locations of a read against the reference. In the second step, this alignment is refined using more accurate methods, such as Smith-Waterman, on the limited subset of locations. To scan the reference or the input reads more quickly, most alignment methods construct auxiliary data structures called *indices*. Based on the type of index used, we can distinguish two main families of short-read alignment methods: hash table-based methods and Burrows-Wheeler transform-based methods (Flicek and Birney 2009).

The hash-based alignment methods build a so-called hash table data structure (Sanders et al. 2019) from the sequence data to index and scan the data. These methods were the first to attempt short-read alignment and are effectively inspired by the BLAST algorithm (see also Section 3.1.5). They hash either the set of input reads or the reference genome (the database), then use the other set (the query) to scan the hash table. More specifically, they analyse the query to find all k-mers (small, fixed length subsequences) and keep the position of each k-mer in a hash table, using the k-mer as the key. Then, the database is scanned for exact matches to the k-mer, called seeds. These seeds are then extended on both sides and Smith-Waterman is used to produce a final alignment. Examples are MAQ (`https://maq.sourceforge.net/`) (reads hashed) (Li, Ruan and Durbin 2008), or SOAP (`http://soap.genomics.org.cn/`) (reference hashed) (Li et al. 2008).

Burrows-Wheeler transform methods use the FM-index data structure proposed by Ferragina and Manzini (2000). This index is a practical improvement on the suffix array: a data structure that stores all the suffixes of a string to allow fast string matching. Ferragina and Manzini showed that a suffix array can be stored more efficiently if generated from the Burrows-Wheeler transformed version of the sequence rather than the original. The FM-index retains the rapid subsequence search of a regular suffix array, but the final index is much smaller. Such Burrows-Wheeler transform implementations are typically much faster than their hash-based counterparts with similar sensitivity levels (e.g. BOWTIE is 30 times faster than MAQ). Examples are Bowtie2 (https://bowtie-bio.sourceforge.net/bowtie2/index.shtml) (Langmead and Salzberg 2012), Maw (https://maq.sourceforge.net/) (Li and Durbin 2009), SOAP2 (http://soap.genomics.org.cn) (Li et al. 2009b).

Once the reads have been aligned to the reference genome, a consensus sequence or variant sequences can be called using specialised tools that take into account both the majority nucleotide for every site as well as the sequence read quality at each base (Li et al. 2009a; Li 2011; Danecek et al. 2021).

3.3.2 *De novo* assembly methods

The *de novo* assembly methods can be divided into those using an overlap-layout-consensus strategy or de Bruijn graphs as underlying data structure (Staden 1979; Idury and Waterman 1995; Compeau, Pevzner and Tesler 2011; Flicek and Birney 2009). Both frameworks require at least a portion of the reads to be longer than the longest near-identical repeat in the genome to create long assemblies *de novo*.

Overlap-layout-consensus methods use overlaps between reads to find the most likely linear consensus of all of them. This read-centric method works best for long-read methods, like Sanger sequencing and third-generation sequencing technologies. It is computationally infeasible for second-generation sequencing data because of the large number of reads and overlaps that need to be tracked.

The second assembly framework uses the de Bruijn graph data structure (de Bruijn 1946; Compeau, Pevzner and Tesler 2011). A de Bruijn graph is based on k-mers: it has a node for every k-mer observed in the sequence set and an edge between nodes if these k-mers are observed next to each other in a read. Reads will be split across their component nodes, and if the sequence contains a repeat region, this will be stored as a set of adjacent k-mers that many reads pass through. On the edges of the repeat, the leading and trailing k-mers will be connected to several different k-mers, representing the distinct positions of this repeat in the genome. How the assembler resolves the resulting forks and bubbles in the de Bruijn graph is one of the main distinguishing features of the method implementations. The main requirement is that the reads have to be longer than the k-mer length, whereby larger k-mers are better but have to be supported by sufficient coverage. The big advantage of this method is that the graph can be constructed in $\mathcal{O}(n)$, where n is the number of reads. However, the memory requirements are often still a limiting factor. Examples are VELVET (https://

github.com/dzerbino/velvet) (Zerbino and Birney 2008), ABySS (https://www.bcgsc.ca/resources/software/abyss) (Simpson et al. 2009), SPAdes (https://cab.spbu.ru/software/spades/) (Bankevich et al. 2012).

4 Genetic associations

Once sequences of different individuals have been aligned using local or global alignment tools, the differences between genotypes can be investigated. One way to compare sequences is to look for *single nucleotide polymorphisms* (SNPs[1]) and identify their phenotypic consequences. A SNP refers to variation at a single position in the DNA sequence among individuals. If more than 1% of a population carries the same nucleotide at a specific position in the DNA sequence (compared to the majority nucleotide at that site), this variation is called a SNP (NIH National Cancer Institute 2023). SNPs can occur both within coding regions and in noncoding regions of DNA. If it occurs in the coding region, we have a connection with the terms introduced earlier: a gene has more than one allele if a SNP occurs within this gene. In such a case, the SNP may lead to variation in the amino acid sequence and, consequently, in the phenotype.

By comparing the characters present at each site, we can identify SNPs in an alignment. Note that if we know which site in the alignment harbours a SNP, we can target this SNP position in the genome directly via microarray genotyping (these technologies have been reviewed in Distefano and Taverna (2011)), which allows quick screening of many individuals for specific SNPs without sequencing. Targeting specific positions like this has the added benefit of not requiring an alignment, saving time and resources.

A key question is whether the different alleles at a SNP position — called *SNP-alleles* — are linked to different phenotypes. For example, the genome of any two people is 99.9% identical (The International HapMap Consortium 2003). Not only do the 0.1% non-identical sites determine physical appearance, but they also impact the risk of developing genome-associated diseases such as Alzheimer's disease (Corder et al. 1993) or type II diabetes (Altshuler et al. 2000). The HapMap consortium was initialised to identify the SNPs in the human genome that might affect human health (The International HapMap Consortium 2003). If we know that certain SNPs are associated with a malignant trait, we can examine the parts in the DNA around these SNPs to identify the gene or genes responsible for the trait, which is a basis for developing treatment options.

This chapter will discuss the case-control setup, a very common analysis tool for studying genetic associations with certain diseases. This analysis tool and its various extensions are commonly referred to as *genome-wide association studies (GWAS)*. Importantly, within the framework of this chapter, we assume that all considered sites (the sites with a SNP) are unlinked, meaning that they have an independent evolutionary history. This may be the case

[1] SNP is pronounced as [snɪp].

if recombination rapidly breaks up any linkage between the considered sites (e.g. in the case of sexual reproduction). In Chapters 8 and 11, we discuss methodology for datasets where full or partial linkage exists between the considered sites.

4.1 Testing for associations

4.1.1 The case-control setup

The case-control setup was developed to identify whether a common genetic variant might be associated with a specific disease. In this type of analysis, a large number of individuals are recruited for a study. This group is then divided into the "case" (diseased) and "control" (healthy) groups. For each individual, the alleles for thousands of SNP positions are determined (most commonly using microarrays and rarely through whole genome sequencing). Microarrays are comparable to sequencing techniques with which it is possible to determine the nucleotide at a given location in the genome (e.g. see the first publication using microarrays by Schena et al. (1995)). The microarrays from Illumina or Affymetrix are most commonly used for GWAS.

In what follows, we assume that only two alleles occur at each SNP position: the major and the minor variant. Each SNP-allele is checked for its association with the disease status by calculating the *odds ratio (OR)*. This is the ratio of the odds of having the disease amongst individuals with the minor variant at the SNP position over the odds of having the disease amongst individuals with the major variant at the SNP position:

$$\text{OR} = \frac{\left(\dfrac{\text{number of } \textbf{diseased} \text{ individuals with } \textbf{minor} \text{ variant at SNP position}}{\text{number of } \textbf{healthy} \text{ individuals with } \textbf{minor} \text{ variant at SNP position}}\right)}{\left(\dfrac{\text{number of } \textbf{diseased} \text{ individuals with } \textbf{major} \text{ variant at SNP position}}{\text{number of } \textbf{healthy} \text{ individuals with } \textbf{major} \text{ variant at SNP position}}\right)}. \tag{4.1}$$

If the odds ratio is greater than one, the minor variant is found more often in the diseased individuals than in the healthy group. On the other hand, if the ratio is smaller than one, the minor variant is present more frequently in the healthy group than in the diseased group. This first hints at whether a minor variant might play a role in the particular disease.

However, to make a statement about our confidence in the SNP-allele playing a role in a specific disease, we need to calculate the p-value. To calculate a p-value, we need to define the *null hypothesis* clearly. The p-value and null hypothesis are generally defined in Box 1 on page 24. Here, the null hypothesis is:

$\mathcal{H}_0$: *The minor variant does not have an effect on the disease.*

Observed	Case	Control	Row sums
Minor variant	49	6	55
Major variant	47	44	91
Column sums	96	50	146

Table 4.1: Contingency table for the age-related macular degeneration (AMD) GWAS.

More precisely, this means that the diseased people are a random subset of the whole population and, therefore, independent of the allele they carry. Thus, the number of cases with the minor allele follows a hypergeometric distribution (Box 9 on page 76).

4.1.2 Calculating the p-value in a GWAS

In this section, we show how to calculate the p-value using the null hypothesis $\mathcal{H}_0$ given above. We do so using data from a very early landmark GWAS published in 2005. It investigated the association of genetic variants with macular degeneration, an age-related eye disease that causes loss of vision (Klein et al. 2005). With the GWAS approach, the authors could identify two SNPs as risk factors for developing this condition.

In this study, 96 individuals suffering from age-related macular degeneration (AMD) (cases) and 50 individuals not suffering from this disease (controls) were enrolled. In total, 116 204 SNP positions were tested per individual, with 103 611 SNP sites of good quality included in the final data analysis.

On SNP rs380390, the minor variant is a C on both alleles. 49 cases expressed a C on both alleles, the other 47 cases expressed a G, or a mix (e.g. one G, one C). 6 controls expressed a C on both alleles, the other 44 controls expressed G, or a mix. These numbers were extracted from Klein et al. (2005, Figure 1B).

These data can be represented in the form of a *contingency table* (see Box 8 on page 75), shown in Table 4.1.

Could SNP rs380390 be associated with AMD? To answer this question, we first calculate the odds ratio, OR, using Equation (4.1):

Expected	Case	Control	Row sums
Minor variant	36.16	18.84	55
Major variant	59.84	31.16	91
Column sums	96	50	146

Table 4.2: Expected contingency table for the age-related macular degeneration (AMD) GWAS.

$$\begin{aligned}\mathrm{OR} &= \frac{\left(\dfrac{\text{number of }\textbf{diseased}\text{ individuals with }\textbf{minor}\text{ variant at SNP position}}{\text{number of }\textbf{healthy}\text{ individuals with }\textbf{minor}\text{ variant at SNP position}}\right)}{\left(\dfrac{\text{number of }\textbf{diseased}\text{ individuals with }\textbf{major}\text{ variant at SNP position}}{\text{number of }\textbf{healthy}\text{ individuals with }\textbf{major}\text{ variant at SNP position}}\right)}\\ &= \frac{49/6}{47/44} = 7.6. \end{aligned} \tag{4.2}$$

Thus, the odds ratio indicates an association between the minor variant and AMD. To calculate the p-value, we apply Pearson's χ^2-test as described in Box 12 on page 79 based on the contingency table. According to Pearson's χ^2-test, we need to calculate the expected number of cases with the minor variant, assuming that this number follows a hypergeometric distribution. Thus, we compute the cell (1,1) of the expected contingency table using the fixed values of the row and column sums (Box 12 on page 79):

$$E_{1,1} = 146 \times \frac{55}{146} \times \frac{96}{146} = 36.16. \tag{4.3}$$

With this entry and the fixed row and column sums, we can complete the expected contingency table shown in Table 4.2.

We now calculate the deviance between the observed and expected numbers using Equation (B12.2) based on unrounded entries of the expected contingency table and obtain $S = 21.34$. As explained in Box 12 on page 79, S is approximately χ^2 distributed (see Box 11 on page 78). This allows us to calculate the p-value as $P(S \geq 21.34) = 3.84 \times 10^{-6}$, which indicates a significant association. However, overall 103 611 SNP positions were considered, and we need to correct for multiple testing in order to make a statistical statement regarding significance.

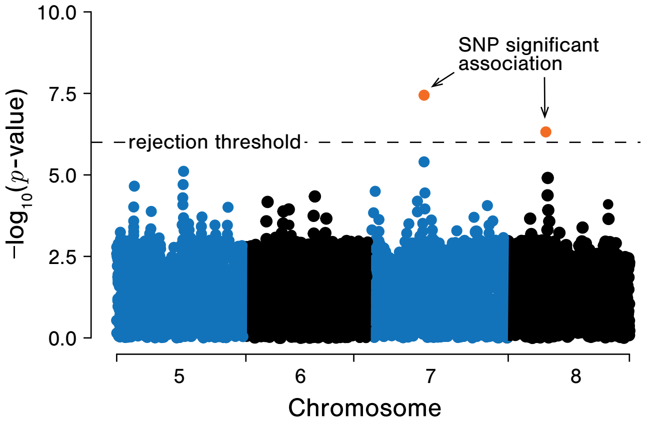


Figure 4.1: Manhattan plot of p-values resulting from a GWAS. For each SNP — ordered on the x-axis according to their position in the genome — the negative logarithm of the p-value from a test of its association with a particular disease is shown on the y-axis. The majority of SNPs have low $-\log(p-$ values). A few SNPs have exceptionally high negative $-\log(p-$values). The $-\log(p-$values) of two SNPs positions lie above the rejection threshold and thus point to a possibly significant association.

4.1.3 Correcting for multiple testing

If many SNP positions are evaluated for their association with the disease status at the same time, the p-values of these tests can be visualised to spot significant trends. One widely used method is the so-called *Manhattan plot*, where the individual p-values are plotted on the y-axis and the position of the SNP in the genome (or on the chromosome) on the x-axis (Figure 4.1). With this method, one can visually identify the chromosome and potentially the gene with the most SNPs associated with a specific disease.

In the following, we discuss how to statistically correct for testing significance in hundreds of thousands of different SNP locations at the same time. As explained in Box 1 on page 24, the p-value of a single statistical test describes how likely it is to obtain the observed outcome or something more extreme, given the null hypothesis. The null hypothesis can be rejected if the p-value is lower than a pre-defined significance level (the rejection threshold). In GWAS, if we were to reject the null hypothesis using the rejection threshold 0.05 for each SNP location, the cumulative probability of the complete study to detect a false positive would be much higher than 0.05, meaning that the significance level α is > 0.05. To counteract this, we need to employ strategies for correcting for multiple testing. One such strategy is the so-called *Bonferroni correction*.

Assume that we test n independent SNP sites. Instead of rejecting if the p-value is < 0.05, we reject the null hypothesis if the p-value $< {}^{0.05}/_{n}$. This rejection threshold guarantees that the significance level is $\alpha = 0.05$, meaning that the cumulative probability of detecting a false positive is smaller than 0.05 (Van den Oord 2008).

In the study on age-related macular degeneration, the authors used the Bonferroni correction (Klein et al. 2005). In total, they included 103 611 SNP sites in the data analysis, and the null hypothesis was rejected when the $p-$value was smaller than the rejection threshold ${}^{0.05}/_{103\,611} = 4.8 \times 10^{-7}$. We obtained a p-value of 3.84×10^{-6}, and thus we do not reject the null hypothesis. However, we note that we performed a rather strict test: we focussed on the association of `C` on both alleles with the disease, while each `C` allele may have an association with the disease. Indeed, when using the allelic counts without insisting on homozygous alleles, the authors obtain a Bonferroni-corrected p-value of 4.2×10^{-8} indicating a significant association of SNP rs380390 with AMD.

4.2 Potentials and drawbacks

The GWAS introduced simultaneous screening of thousands of genetic variants for their association with the disease. As of March 2024, the GWAS Catalog (`https://www.ebi.ac.uk/gwas/`) contains 6 779 publications with 580 440 identified unique SNP-trait associations (Sollis et al. 2022; Grimm et al. 2017). The SNPedia (`https://www.snpedia.com/index.php/SNPedia`) is an attempt to collect all relevant SNPs in the human genome with the associated disease risk (Cariaso and Lennon 2012). We end this chapter by mentioning the limitations of the described setup and how to overcome them.

The case-control setup makes sense when one can clearly distinguish between the case and control group. However, some diseases range from expressing mild to very severe symptoms. This information on the quantitative trait of disease severity can be used to perform an analysis of variance (ANOVA). In this case, the null hypothesis is that there is no difference between the phenotypic means of any genotype class (Bush and Moore 2012).

While GWAS uncovers the association of SNP-alleles with phenotypes, it cannot uncover causation. This means that certain allele patterns at a SNP position may be associated with a particular disease but are not the cause. Further molecular biology experiments are needed to show if a significant SNP is indeed responsible for the disease (causation) and how it contributes to the disease status (mechanistic understanding).

By using a GWAS, we assume that variation in different SNP positions is independent of each other, implying we assume that there is no linkage between sites. Biologically, linkage is broken up quickly between sites if there is a lot of recombination and sites of interest are far apart. However, in reality, samples intended for GWAS may show dependencies due to population structure or also some remaining linkage across the genome. GWAS methods have been introduced that take into account such dependencies, for example, by using linear mixed models or introducing principal components as covariates (Zhou and Stephens 2012). Correcting for dependencies between samples becomes especially important for microbial

Box 8: Contingency table test

Contingency table tests are statistical tests that allow us to test for an association between two or more classes with two or more characteristics each. The observations are represented in a *contingency table*. We consider here a contingency table for two classes, X and Y, with two characteristics each: X_1, X_2 in class X, and Y_1, Y_2 in class Y. In total, there are n observations that can fall into any of the four categories, with the results summarised in the following contingency table:

	Y_1	Y_2	Row sums
X_1	a	b	$a+b$
X_2	c	d	$c+d$
Column sums	$a+c$	$b+d$	n

where $n = a+b+c+d$. Entry (i,j) in the matrix describes how many observations showed X_i and Y_j.

Contingency table tests are designed to test whether a characteristic within X is associated with a characteristic within Y. We use two examples of these tests. For small datasets, we use *Fisher's exact test* (see Box 10 on page 77; applied to data in Chapter 8); for large datasets, this test is computationally infeasible. In such situations, we use *Pearson's* χ^2*-test* (explained in Box 12 on page 79; applied to data in Chapter 4).

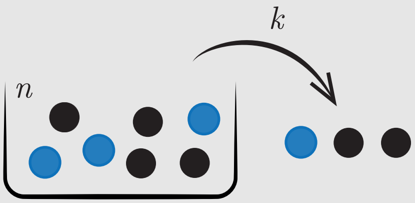
GWAS (when considering bacteria that reproduce clonally or viruses rather than eukaryotes): phylogenetic and clustering approaches are common options to account for these dependencies (Power, Parkhill and de Oliveira 2017). Tools specifically developed to address these challenges (Power, Parkhill and de Oliveira 2017; San et al. 2020) have been successfully applied, for example, to identify antibiotic resistance variants in bacteria (Ma et al. 2020; The CRyPTIC Consortium 2022).

To learn more about GWAS and its extensions, please refer to Pearson and Manolio (2008), Bush and Moore (2012), Scherer and Visscher (2016), Tam et al. (2019) and Uffelmann et al. (2021).

In the following chapters, we will discuss how to deal with sequence information in case of strong linkage and investigate associations between genotypes and phenotypes under strong linkage in Chapter 8. In Chapter 11, we will outline the first advances in the field when sites have intermediate amounts of linkage.

Box 9: Hypergeometric distribution

The *hypergeometric distribution* is a discrete probability distribution that describes the probability of drawing i balls of one type in k draws without replacement from an urn containing n balls of two types. The urn contains r blue and s black balls, $n = r + s$. In the sketch below, the urn has $n = 7$ balls, of which $r = 3$ balls are blue and $s = 4$ balls are black; we draw $k = 3$ times, and $i = 1$ balls are coloured blue.



We assign a capital letter for the random variable "the number of blue balls amongst k drawn balls", say R_k. What is the probability of drawing exactly i blue balls? In general, we can calculate this probability as the number of possible draws of k balls yielding exactly i blue balls divided by the total number of possible k ball draws from the urn with n balls.

There are $\binom{r}{i}$ possibilities (see Box 2 on page 25) to obtain i blue balls, and $\binom{s}{k-i}$ possibilities to draw $k - i$ black balls from the urn. Further, there are in total $\binom{n}{k}$ ways to draw k balls out of the urn without replacement. Thus, the probability to draw i blue balls amongst the k drawn balls is

$$P(R_k = i) = \frac{\binom{r}{i}\binom{s}{k-i}}{\binom{n}{k}}. \tag{B9.1}$$

The mean of this hypergeometrically distributed random variable is

$$\mathrm{E}(R_k) = \frac{kr}{n}, \tag{B9.2}$$

and its variance is

$$\mathrm{Var}(R_k) = \frac{kr(n-r)(n-k)}{n^2(n-1)}. \tag{B9.3}$$

The hypergeometric distribution has many applications and is used in this book in Fisher's exact test (Box 8 on page 75) and for comparative methods (Section 8.1).

Box 10: Fisher's exact test

Fisher's exact test|textbf was developed by the British mathematician Ronald Fisher (1890-1962) to test the claims of a British lady who said that she could distinguish between two modes of preparing a British tea (which is always drunk with milk). She claimed it was possible to taste the difference between whether one adds the tea to the cup and then milk (TIF = tea into cup first) or vice versa (MIF = milk into cup first). This example was published under "Mathematics of a Lady Tasting Tea" (Fisher 1956).

The two classes in this example are how the tea was prepared (with the characteristics milk first versus tea first) and the prediction of the lady tasting the tea (with the characteristics predicted milk first and predicted tea first). Thus, the null hypothesis is:

$\mathcal{H}_0$: *The mode of preparing the tea and the lady's predictions are independent.*

The observations from Fisher's tea example can be written in a contingency table, where X stands for the mode of preparing tea, $X_1 =$ TIF, and $X_2 =$ MIF. Y represents whether the lady predicts tea in cup first, Y_1, or milk first, Y_2. If the mode of preparing tea and the lady's observation were independent, this experiment is analogue to the urn experiment described in Box 9 on page 76. The number of blue balls corresponds to the cups of tea prepared TIF, $a + b$, and the number of black balls corresponds to MIF, $c + d$. The number of times the lady predicts tea in cup first (Y_1) when this is, in fact, correct (X_1) is a random variable Z. We refer to a specific realisation of this random variable with a. Then, we can rewrite our null hypothesis:

$\mathcal{H}_0$: *The random variable Z follows a* hypergeometric distribution.

From Equation (B9.1), we write:

$$P(Z = a) = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}. \tag{B10.1}$$

As explained in Box 1 on page 24, the p-value is calculated by summing the probabilities to obtain the observed or more extreme results. In the tea example, a more extreme result would be obtained if higher predictions of correct tea preparation were made, that is, if the entry $(1, 1)$ in the table was greater than a. Thus, the p-value is

$$p = \sum_{i=a}^{a+b} \frac{\binom{a+b}{i}\binom{c+d}{a+c-i}}{\binom{n}{a+c}}. \tag{B10.2}$$

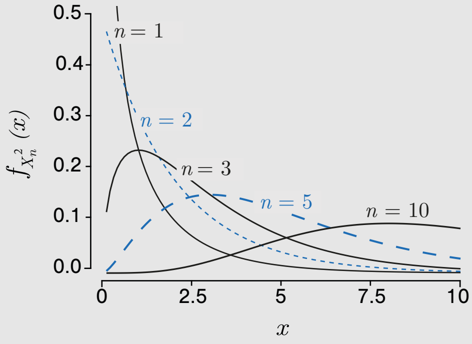
This calculation can be done by hand for small numbers (in the tea example, only eight cups of tea were brewed, 4 with TIF and 4 with MIF). We will see an example of this procedure in Chapter 8. The summation is computationally infeasible for larger datasets. As a side note, the lady supposedly could determine all cups of tea correctly (Salsburg 2002).

Box 11: χ^2 distribution

The χ^2 *distribution (chi-squared distribution)*[2] plays an important role in statistics, more precisely in statistical testing. This distribution is obtained through the convolution of distributions of independent and identically distributed random variables Y_i that are normally distributed with mean μ and variance σ^2 (see Box 13 on page 80), in mathematical notation $Y_i \sim \text{Normal}(\mu, \sigma^2)$, in the following way:

$$X_n^2 = \sum_{i=1}^{n} \frac{(Y_i - \mu)^2}{\sigma^2}. \tag{B11.1}$$

The distribution of X_n^2 is called χ^2 distribution with n degrees of freedom (in short χ_n^2) and takes different shapes, as shown in the following figure.



Its density function is

$$f_{X_n^2}(x; n) = \frac{1}{2^{n/2}\Gamma(n/2)x^{\frac{n}{2}-1}e^{-\frac{x}{2}}}. \tag{B11.2}$$

The Γ function is further explained in Equation (B14.2) in Box 14 on page 81.

The mean of this distribution is

$$\text{E}(X_n^2) = n, \tag{B11.3}$$

and its variance is

$$\text{Var}(X_n^2) = 2n. \tag{B11.4}$$

The χ_n^2 distribution approximates many distributions that appear in statistical testing. This means that if the exact distributions are not known or cannot be calculated, one can assume a χ_n^2 distribution when determining the p-value (see also Box 1 on page 24). For an outcome x where $f_{X_n^2}(X \geq x; n) = \alpha$, we also write $x = \chi_{n,\alpha}^2$. The p-value of x under a χ^2 distribution can be found in χ^2-tables or can be directly computed using tools such as R (`https://www.r-project.org/`).

We will see applications of the χ^2 distribution when calculating p-values in Box 8 on page 75 and more generally in Chapter 7. More information on the χ^2 distribution can be found in Sokal and Rohlf (2012).

[2] χ is pronounced as [kaɪ].

Box 12: Pearson's χ^2-test

In Box 10 on page 77, the p-value for data in a contingency table was calculated using Fisher's exact test. The p-value calculation becomes computationally infeasible for larger datasets. A way to calculate the p-value for large datasets is to use *Pearson's* χ^2*-test* (Pearson 1900). In fact, it only works for reasonably large datasets. The term χ^2*-test* is used for all statistical tests where the distribution of interest can be approximated by a χ^2 distribution under the null hypothesis.

When doing an experiment, we fill in the contingency table entries $\mathrm{O}_{i,j}$ according to the observations and also calculate the row and column sums as presented in Box 8 on page 75. Now, we further fill in a second contingency table with the expected value $\mathrm{E}_{1,1}$ under the hypergeometric distribution. This can be easily calculated by

$$\mathrm{E}_{1,1} = n \times \frac{a+b}{n} \times \frac{a+c}{n}. \tag{B12.1}$$

Given this value, we can fill in the remaining entries of the table as the row and column sums are fixed.

We define the following data transformation:

$$\mathrm{S} = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}. \tag{B12.2}$$

This sum describes the average deviance between the observed and the expected data, given that the null hypothesis is true. One can demonstrate that S is approximately χ_1^2 distributed, meaning that it has a χ^2 distribution with one degree of freedom (Fisher 1922; Chernoff and Lehmann 1954). According to the definition of the p-value as the probability of obtaining the observed result, s, or a more extreme result (see Box 1 on page 24), we can approximate the p-value using the χ_1^2 distribution by

$$p\text{-value} = P_{\chi_1^2}(\mathrm{S} \geq s). \tag{B12.3}$$

As explained in Box 1 on page 24, we reject the null hypothesis if the p-value is less or equal to the pre-defined rejection threshold.

Box 13: Normal distribution

The *normal distribution*, also known as the *Gaussian distribution* or colloquially as the *bell curve*, is a probability density defined on the continuous values of a single random variable X. Its probability density function is

$$f_X(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}. \tag{B13.1}$$

It has two parameters, μ and σ, which are the mean and standard deviation for this distribution, respectively:

$$\mathrm{E}(X) = \mu, \tag{B13.2}$$

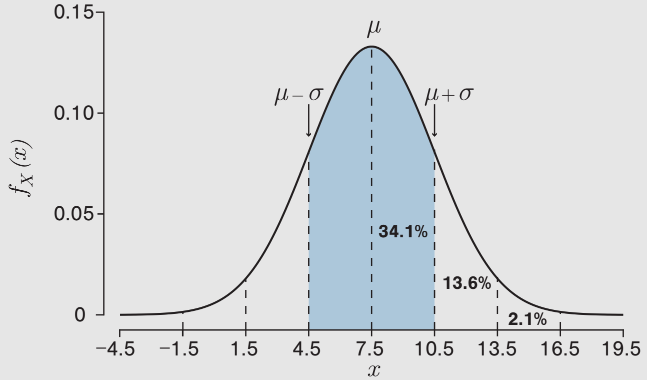
$$\mathrm{Var}(X) = \sigma^2. \tag{B13.3}$$

Writing $X \sim \mathrm{Normal}(\mu, \sigma^2)$ means that random variable X is normally distributed with mean μ and standard deviation σ (variance σ^2).

Simple properties include:

1. if $X \sim \mathrm{Normal}(\mu, \sigma^2)$ and $Y = X + c$, then $Y \sim \mathrm{Normal}(\mu + c, \sigma^2)$;
2. if $X \sim \mathrm{Normal}(\mu, \sigma^2)$ and $Z = {}^{(X-\mu)}/_{\sigma}$, then $Z \sim \mathrm{Normal}(1, 0)$;
3. if $X \sim \mathrm{Normal}(\mu_X, \sigma_x^2)$ and $Y \sim \mathrm{Normal}(\mu_Y, \sigma_Y^2)$, then their sum $Z = X + Y$ is a random variable with $Z \sim \mathrm{Normal}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$.

The following figure displays the probability density function for a normal distribution with $\mu = 7.5$ and $\sigma = 3$:



The normal distribution is ubiquitous in statistics, principally as the result of the *central limit theorem*. Informally, this theorem implies that if we define the random variable Z_n to be the average of n independent samples from any single distribution with mean m and finite variance s^2, the distribution for Z asymptotes to a normal distribution centred on m and with variance ${}^{s^2}/_{n}$ as the sample count n becomes large.

Box 14: Γ distribution (gamma distribution)

The Γ *distribution* is defined on $[0, \infty)$ with parameters $\alpha > 0$ (shape), and $\beta > 0$ (rate), and has the probability density function

$$f_X(x; \alpha, \beta) = \frac{\beta^\alpha}{\boldsymbol{\Gamma}(\alpha)} e^{-\beta x} x^{\alpha-1}, \tag{B14.1}$$

where $x \geq 0$ and

$$\boldsymbol{\Gamma}(\alpha) = \int_0^\infty e^{-t} t^{\alpha-1}\, \mathrm{d}t. \tag{B14.2}$$

The latter function is also called Γ function and fulfils $\boldsymbol{\Gamma}(n) = (n-1)!$ for natural numbers $n \in \{1, 2, 3, \ldots\}$.

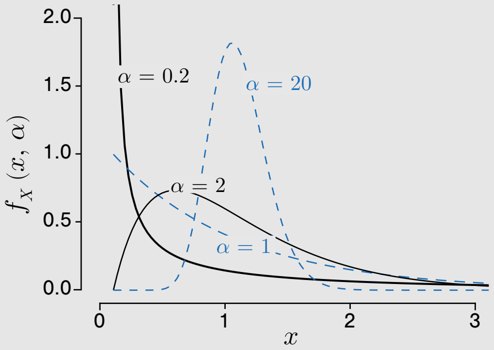
A Γ distributed random variable X with parameters α, β ($X \sim \Gamma(\alpha, \beta)$) has the mean

$$\mathrm{E}(X) = \frac{\alpha}{\beta}, \tag{B14.3}$$

and variance

$$\mathrm{Var}(X) = \frac{\alpha}{\beta^2}. \tag{B14.4}$$

In this book, we will mostly use the Γ distribution with $\alpha = \beta$ (that is, with mean 1). This distribution is quite flexible with respect to different values of α, as shown in the following plot.



5 Molecular evolution

In this chapter, we introduce evolutionary models describing the change of genetic sequences through time, resulting in molecular evolution. Molecular evolution models allow us to quantify evolutionary processes acting on the sequences, for example, the rates of substitution of one base by another. Furthermore, they are an essential component of phylogenetic reconstruction methods. Evolutionary models were designed for three levels of evolution: the level of DNA or RNA sequences (genotypic level), the level of codons (triplets of nucleotides encoding an amino acid), and the level of amino acid sequences (phenotypic level).

The models presented here (those commonly used) only account for point mutations but not for insertions/deletions (indels), inversions, or recombination. When doing downstream analysis, these models should be used only on parts of the multiple sequence alignment (MSA) that differ primarily (or only) due to point mutations. In practice, the MSAs often have some gaps indicating indels, but these gaps are commonly treated as unknown nucleotides (this treatment also applies to the models discussed below). Investigation is ongoing to assess the potential biases stemming from these assumptions.

We will now first discuss the difference between mutations and substitutions, then define the commonly used sequence evolution models at the DNA or RNA level, followed by a discussion of the general properties of these models. Last, we extend this framework to study the evolution of sequences at the codon and the amino acid level.

5.1 Substitutions vs. mutations

Although point mutations occur during DNA replication (as discussed in the introduction, Chapter 1), molecular evolution models do not model the process of replication directly. They all assume that character changes occur at each site within a sequence at some rate through time. The reasoning for this modelling choice becomes clear when considering the evolution of species. Individuals of a species have identical characters at most positions in their sequence (for example, only 0.1% of the positions vary within the human population). A change at a particular non-variant site happens at the species level if a mutation occurs in the germ line of an individual, and this mutated cell gives rise to offspring. This offspring carries the mutation both in the somatic and germ cells. Eventually, by chance or selection (see Section 1.2), the individual with the mutated genome may spread in the population until its offspring, in turn, make up the vast majority of the population, and we say that the mutation became fixed. A fixed mutation is called a *substitution*. Thus, substitutions occur at some stage during a species' existence but not at speciation. When considering a different biological unit, for example, an infected host, we observe a similar process: one pathogen (e.g. one bacterium or one virion) in one infected individual mutates, and the mutation may be fixed within the pathogen population in this particular infected individual.

Based on these considerations, models for sequence evolution refer to changes in the characters as *substitutions*, and the models themselves are often called *substitution models*.

We highlight that when using molecular evolution models for questions where biological units are single cells or virions, the changes in the sequences are indeed mutations, but our models still call them substitutions. One may argue that we should model this mutation process by allowing changes only at replication events instead of at any time. This means we should not use the substitution models presented here; instead, we should use models linking replication with mutation. However, we can reason that these substitution models are still appropriate even in such scenarios by imagining that we track a single cell or virion through time. Upon replication, we follow one of its offspring. When this offspring replicates, we again follow one of its offspring, and so on. The tracked lineage accumulates mutations through time (happening at the replication events), and given that we have many replications, each with a few mutations, we may choose to approximate the sequence changes by a model where changes accumulate at some rate (as assumed in the substitution models below). In doing so, we implicitly assume that we sampled only a few individuals from the population. This sparse sampling implies that the observed mutations typically occur along lineages due to unsampled replication events rather than at the branching of two observed individuals. If the assumption of sparse sampling is violated, we would need new models that only allow mutations at replication.

5.2 General theory on nucleotide substitution models

We will now introduce the general structure of nucleotide substitution models. They are all specified by a substitution rate matrix (Section 5.2.1), which determines transition probabilities between nucleotides (Section 5.2.2) and gives rise to a Markov chain model for sequence evolution (Section 5.2.4).

5.2.1 Substitution rate matrix

Each site in a DNA sequence can have one of the four nucleotides A, C, T, or G. A site in a sequence with a particular nucleotide may change through time to a different nucleotide. Common evolutionary models assume that the change from state i into state j (where $i, j \in \{\mathtt{T}, \mathtt{C}, \mathtt{A}, \mathtt{G}\}$ and $i \neq j$) happens in an infinitesimally small time interval Δt with probability $q_{i,j}\Delta t$. $q_{i,j}$ is called the *substitution rate* from i to j. Note that per this definition, the probability of a change from i to j is the same at any point in time.

"Infinitesimally" small here means that the time step is so small that either nothing happens or only one event occurs with a non-negligible probability. This implies that several sequential changes (such as a change from i to k to j) have a negligible probability of happening within this time step. We can mathematically specify the infinitesimally small probability of more than one event. Let us denote the rate of any substitution with q. Two events happen with probability $(q\Delta t)^2$, three events with $(q\Delta t)^3$, and so on. In summary, the probability of more than one event is of the order of $\mathcal{O}(\Delta t^2)$. This term quickly goes to zero for small Δt. Below, we derive the equations acknowledging terms of order $\mathcal{O}(\Delta t^2)$ and show how they disappear when taking the limit $\Delta t \to 0$.

The most convenient way to denote the substitution rates is in a matrix, where the rows denote the original state and the columns denote the substitution, referred to as *substitution rate matrix Q*:

$$Q = \begin{array}{c} \\ \mathsf{T} \\ \mathsf{C} \\ \mathsf{A} \\ \mathsf{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathsf{T} & \mathsf{C} & \mathsf{A} & \mathsf{G} \end{array} \\ \begin{pmatrix} \cdot & a & b & c \\ d & \cdot & e & f \\ g & h & \cdot & i \\ j & k & l & \cdot \end{pmatrix} \end{array}. \tag{5.1}$$

Note that the order of the nucleotides in the nucleotide substitution rate matrix is not unambiguously defined throughout the literature, so in some sources, the nucleotides are ordered alphabetically. The rates can be read off the matrix in a row-to-column way. So, if we need the G-to-A substitution rate, we will look at the entry in the last row and the third column in our example rate matrix above, corresponding to the rate l.

Although the substitution rates are defined only for the off-diagonal entries, for mathematical convenience, the diagonal entries are set such that each row sums up to zero.

In summary, we can write down the complete substitution rate matrix as follows:

$$Q = \begin{array}{c} \\ \mathsf{T} \\ \mathsf{C} \\ \mathsf{A} \\ \mathsf{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathsf{T} & \mathsf{C} & \mathsf{A} & \mathsf{G} \end{array} \\ \begin{pmatrix} -(a+b+c) & a & b & c \\ d & -(d+e+f) & e & f \\ g & h & -(g+h+i) & i \\ j & k & l & -(j+k+l) \end{pmatrix} \end{array}. \tag{5.2}$$

5.2.2 Transition probability matrix

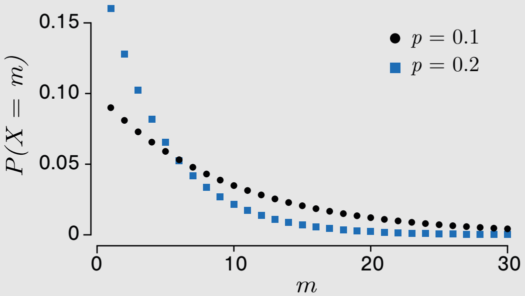
Typically, we are interested in the probability of a nucleotide changing from i to j in a time interval t. This probability in matrix form is $Qt + \mathcal{O}(t^2)$ for small t. In what follows, we derive the probability of a nucleotide changing from i to j in any time interval t; these probabilities are summarised in the *transition probability matrix*.

Box 15: Geometric distribution

Suppose we repeat Bernoulli trials with success probability p until we observe the first success. Defining X as the random variable representing the repetition during which the first success occurs, the probability of X being equal to m is given by the *geometric distribution*,

$$P(X = m) = (1 - p)^{m-1}p. \tag{B15.1}$$

This is the product of the probability for the first $m - 1$ failures with the probability of the success in repetition m. The shape of the geometric distribution for success probabilities 0.1 and 0.2 is shown below.



The mean of the distribution is the inverse of the success probability:

$$\mathrm{E}(X) = {}^{1}\!/_{p}. \tag{B15.2}$$

Note that the process behind the geometric distribution is discrete and memoryless (see Box 24 on page 98 for the definition).

5.2.2.1 Time to first event

We now calculate the time to the first substitution event, which will facilitate the transition probability matrix calculation. We consider a process that generates an event E at rate α. This means that the probability that E occurs once in an infinitesimally small interval of time Δt is

$$P(E \text{ occurs once in } \Delta t) = \alpha\Delta t. \tag{5.3}$$

The probability of two or more events happening in Δt is infinitesimally small. If we assume that this probability is 0, then for some fixed Δt, the probability of an event in each interval is $\alpha\Delta t$, and the probability of no event is $1 - \alpha\Delta t$. The probability of no event in the first $m - 1$ intervals and of the first event in the mth interval is thus $(1 - \alpha\Delta t)^{m-1}(\alpha\Delta t)$. This follows the geometric distribution (Box 15 on page 86) with parameter $\alpha\Delta t$.

In what follows, we remove the approximation that the infinitesimally small probabilities are 0. The probability that E occurs more than once is $\mathcal{O}(\Delta t^2)$ (the probability of two events is $(\alpha\Delta t)^2$, three events $(\alpha\Delta t)^3$, and so on). Let us denote the time until an event happens as X. We can calculate the probability of no event happening within Δt as

$$P(X > \Delta t) = 1 - \alpha\Delta t + \mathcal{O}(\Delta t^2). \tag{5.4}$$

We now aim to determine the probability density of X. We look at a longer time interval τ and divide it into smaller time intervals Δt such that $\tau = k\Delta t$. Then, using the binomial theorem, we obtain:

$$P(X > \tau) = (1 - \alpha\Delta t + \mathcal{O}(\Delta t^2))^k = (1 - \alpha\Delta t)^k + \mathcal{O}(\Delta t^2) = (1 - \alpha\Delta t)^{\tau/\Delta t} + \mathcal{O}(\Delta t^2) \xrightarrow[\Delta t \to 0]{} e^{-\alpha\tau}. \tag{5.5}$$

The limit for $\Delta t \to 0$ in the latter equation holds because the exponential function is $e^x = \lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^n$ (see Box 16 on page 88).

This directly leads to

$$P(0 \leq X \leq \tau) = 1 - e^{-\alpha\tau}, \tag{5.6}$$

which is called the cumulative distribution function of X. The probability density function can then be obtained by differentiating the cumulative distribution function:

$$f_X(x) = \frac{\mathrm{d}P}{\mathrm{d}t}(x) = \alpha e^{-\alpha x}. \tag{5.7}$$

This is the probability density function of an exponential distribution with parameter α (see Box 17 on page 89). This means that an event that occurs at rate α has exponentially distributed waiting times with parameter α.

In summary, we derived the waiting time until the first event. When we made the approximation of the infinitesimally small elements being equal to 0 (discrete time steps of size Δt), we obtained the geometric distribution; when letting Δt go to 0 (assuming continuous time), we obtained the exponential distribution. In Box 18 on page 90, we show that the exponential distribution is a continuous-time limit of the geometric distribution.

5.2.2.2 Transition probability matrix at time 0 and for small time steps

The transition probability for a change from nucleotide i to j in time interval t is notated with $p_{i,j}(t)$. We summarise these probabilities in the *transition probability matrix*:

$$P(t) = (p_{i,j}(t))_{i,j\in\{\mathsf{C},\mathsf{T},\mathsf{A},\mathsf{G}\}}\,. \tag{5.8}$$

To derive the transition probability matrix for any time step t, we first derive the transition probability matrix $P(0)$ and for infinitesimally small time steps Δt. We employ the same properties of rates and probabilities as in the previous sections but use matrix notation.

Box 16: Exponential function

The exponential function can be defined in multiple equivalent ways. In this book, we will use the following definition:

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}. \tag{B16.1}$$

We will now show that

$$e^x = \lim_{n\to\infty} \left(1+\frac{x}{n}\right)^n. \tag{B16.2}$$

Proof: We define $s_n = \sum_{k=0}^{n} \frac{x^k}{k!}$ and $t_n = \left(1+\frac{x}{n}\right)^n$ for any $x \geq 0$ and an integer n. Using the binomial theorem, we can write down t_n:

$$\begin{aligned} t_n &= \left(1+\frac{x}{n}\right)^n = \sum_{k=0}^{n} \binom{n}{k} \left(\frac{x}{n}\right)^k = 1 + x + \sum_{k=2}^{n} \frac{x^k}{k!} \frac{n(n-1)\ldots(n-(k-1))}{n^k} \\ &= 1 + x + \frac{x^2}{2!}\left(1-\frac{1}{n}\right) + \ldots + \frac{x^n}{n!}\left(1-\frac{1}{n}\right)\ldots\left(1-\frac{n-1}{n}\right) \\ &\leq 1 + x + \frac{x^2}{2!} + \ldots + \frac{x^n}{n!} = s_n. \end{aligned} \tag{B16.3}$$

If we let $n \to \infty$, we can conclude from Equation (B16.1) that

$$\limsup_{n\to\infty} t_n \leq \lim_{n\to\infty} s_n = e^x. \tag{B16.4}$$

We need to use lim sup in Equation (B16.4), because we do not yet know whether the term $1 + x + \frac{x^2}{2!}\left(1-\frac{1}{n}\right) + \ldots + \frac{x^n}{n!}\left(1-\frac{1}{n}\right)\ldots\left(1-\frac{n-1}{n}\right)$ converges.

For all $2 \leq m \leq n$ we can state that

$$1 + x + \frac{x^2}{2!}\left(1-\frac{1}{n}\right) + \ldots + \frac{x^m}{m!}\left(1-\frac{m-1}{n}\right) \leq t_n, \tag{B16.5}$$

because we consider only the first $m+1$ terms of t_n on the left side of the equation. Taking the limit $n \to \infty$ on both sides of this equation, we obtain:

$$s_m = 1 + x + \frac{x^2}{2!} + \ldots + \frac{x^m}{m!} \leq \liminf_{n\to\infty} t_n. \tag{B16.6}$$

Again, we need to take the lim inf because we do not know whether t_n converges. As the right side of this equation is true for all $m \leq n$, we can now take the limit $m \to \infty$ on the left side and obtain

$$e^x = \lim_{m\to\infty} s_m \leq \liminf_{n\to\infty} t_n. \tag{B16.7}$$

Combining Equation (B16.4) and Equation (B16.7), we obtain,

$$\limsup_{n\to\infty} t_n \leq e^x \leq \liminf_{n\to\infty} t_n. \tag{B16.8}$$

Thus, the limit of t_n exists and is equal to e^x, proving Equation (B16.2). □

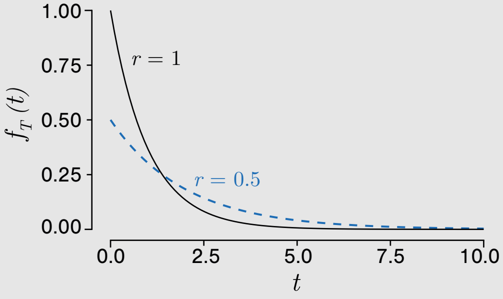
This proof is an extension of the proof for $x = 1$ in Rudin (1976).

Box 17: Exponential distribution

The *exponential distribution* is defined for a continuous positive random variable T. Often, this variable represents a waiting time until some event occurs. The exponential distribution takes its name from the exponential function (Box 16 on page 88) appearing in its probability density function:

$$f_T(t;r) = re^{-rt}. \tag{B17.1}$$

This function can be interpreted as the product of the probability e^{-rt} that the event does not occur in the interval before time t with the probability density r that it occurs immediately after this interval. Its single parameter r is the rate of the exponential distribution. The shape of the probability density function for rates 0.5 and 1 is shown below.



The mean of this distribution is

$$\mathrm{E}(T) = {}^1\!/_r. \tag{B17.2}$$

The exponential distribution can be seen as the continuous analogue of the geometric distribution (see also Box 18 on page 90). While the geometric distribution describes the number of trials (with success having probability p and failure having probability $1 - p$) before success, the exponential distribution describes the time before an event with a fixed rate r.

An important property is the following. If T_i are exponentially distributed random variables with rates r_i, for $i \in [1, \ldots, M]$, and we define X as the minimum of these variables, X itself is exponentially distributed with the rate $R = \sum_{i=1}^{M} r_i$.

We can easily show this for the two variable ($M = 2$) case:

$$\begin{aligned} P(X = x) &= P(T_2 > T_1, T_1 = x) + P(T_1 > T_2, T_2 = x) \\ &= P(T_2 > T_1 | T_1 = x)P(T_1 = x) + P(T_1 > T_2 | T_2 = x)P(T_2 = x) \\ &= e^{-r_2 x}e^{-r_1 x}r_1 + e^{-r_1 x}e^{-r_2 x}r_2 \\ &= e^{-(r_1+r_2)x}(r_1 + r_2). \end{aligned} \tag{B17.3}$$

The generalisation for $M > 2$ is straightforward.

Box 18: Connection between exponential and geometric distribution

We demonstrate that the exponential distribution arises as a certain limit of the geometric distribution. Consider the geometric distribution defined in Box 15 on page 86:

$$P(m|p) = (1-p)^{m-1}p. \tag{B18.1}$$

To connect to the exponential distribution, we first need to map the discrete variable m from the geometric distribution to a continuous time variable. We accomplish this mapping by assuming that the discrete values m correspond to regularly spaced times on a grid where the difference between adjacent times is Δt. Then, we define the time variable as $t = m\Delta t$. Furthermore, we define the rate parameter $r = {}^{p}/_{\Delta t}$, which changes the probability p of success into a measure of success probability per unit time.

The geometric probability distribution function above then becomes the probability for the continuous variable t falling in a particular interval centred on $t = m\Delta t$:

$$P(T \in [t - {}^{\Delta t}/_{2}, t + {}^{\Delta t}/_{2}]|r) = (1-\Delta t r)^{\frac{t}{\Delta t}-1}\Delta t r. \tag{B18.2}$$

The probability density at t is obtained by dividing by Δt and taking the limit $\Delta t \to 0$:

$$\begin{aligned} f(t|r) &= \lim_{\Delta t \to 0} \frac{P(T \in [t - {}^{\Delta t}/_{2}, t + {}^{\Delta t}/_{2}]|r)}{\Delta t} \\ &= \lim_{\Delta t \to 0} (1-\Delta t r)^{\frac{t}{\Delta t}-1} r \\ &= r \lim_{\Delta t \to 0} \exp\left(\left(\frac{t}{\Delta t} - 1\right)\log(1-\Delta t r)\right). \end{aligned} \tag{B18.3}$$

Since $\exp(\cdot)$ is a continuous function, we can bring the limit inside the exponential:

$$\begin{aligned} f(t|r) &= r\exp\left(\lim_{\Delta t \to 0}\left(\frac{t}{\Delta t} - 1\right)\log(1-\Delta t r)\right) \\ &= r\exp\left(\lim_{\Delta t \to 0}\frac{\log(1-\Delta t r)}{\frac{\Delta t}{t-\Delta t}}\right). \end{aligned} \tag{B18.4}$$

Since both the numerator and denominator of $\frac{\log(1-\Delta t r)}{\frac{\Delta t}{t-\Delta t}}$ approach zero as Δt goes to zero, we can apply L'Hôpital's rule to find:

$$\begin{aligned} f(t|r) &= r\exp\left(\lim_{\Delta t \to 0}\frac{\frac{-r}{1-\Delta t r}}{\frac{1}{t-\Delta t} + \frac{\Delta t}{(1-\Delta t)^2}}\right) \\ &= r\exp(-rt). \end{aligned} \tag{B18.5}$$

Thus, as the time between possible outcomes and the probability of individual outcomes of a geometrically distributed random variable becomes small, its probability density distribution approaches that of an exponentially distributed random variable.

When no time has passed ($t = 0$), the probability of a substitution happening is 0, and the probability of staying in the same state is 1. Thus,

$$P(0) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = I. \tag{5.9}$$

The matrix I with 1 on the main diagonal and 0 elsewhere is called the *identity matrix*.

After an infinitesimally small time step Δt, the entry $p_{i,j}(\Delta t)$ is:

$$p_{i,j}(\Delta t) = \begin{cases} q_{i,j}\Delta t + \mathcal{O}(\Delta t^2) & \text{if } i \neq j; \\ 1 - \sum_{k \neq i} q_{i,k}\Delta t + \mathcal{O}(\Delta t^2) & \text{if } i = j. \end{cases} \tag{5.10}$$

That is when $i \neq j$, $p_{i,j}(\Delta t)$ represents the probability of a transition from i to j. In the limit of small Δt, this is simply the instantaneous transition rate $q_{i,j}$ multiplied by the time interval Δt. On the other hand, when $i = j$, $p_{i,i}(\Delta t)$, the probability of no change occurring in the time interval can be easily expressed as 1 minus the probability of any state change occurring:

$$p_{i,i}(\Delta t) = 1 - \sum_{k \neq i} p_{i,k}(\Delta t) = 1 - \sum_{k \neq i} q_{i,k}\Delta t + \mathcal{O}(\Delta t^2). \tag{5.11}$$

By employing $q_{i,i} = -\sum_{k \neq i} q_{i,k}$, we can write:

$$p_{i,i}(\Delta t) = 1 + q_{i,i}\Delta t + \mathcal{O}(\Delta t^2), \tag{5.12}$$

which in matrix notation simplifies to:

$$P(\Delta t) = I + Q\Delta t + \mathcal{O}(\Delta t^2). \tag{5.13}$$

Given the nucleotide substitution rate matrix defined in Equation (5.2), the corresponding transition probability matrix for a small time step Δt is:

$$P(\Delta t) = \begin{pmatrix} 1-(a+b+c)\Delta t & a\Delta t & b\Delta t & c\Delta t \\ d\Delta t & 1-(d+e+f)\Delta t & e\Delta t & f\Delta t \\ g\Delta t & h\Delta t & 1-(g+h+i)\Delta t & i\Delta t \\ j\Delta t & k\Delta t & l\Delta t & 1-(j+k+l)\Delta t \end{pmatrix} + \mathcal{O}(\Delta t^2). \tag{5.14}$$

Note that if we ignore the $\mathcal{O}(\Delta t^2)$ terms, the rows of the transition probability matrix sum up to 1. Indeed, a row of the transition probability matrix describes the probabilities of a nucleotide change to a different one (off-diagonal entries) and the probability of the nucleotide staying the same (diagonal entry). Thus, the sum of the non-negligible probabilities of one event (nucleotide change) or no event (no change) is indeed 1.

5.2.2.3 Transition probability matrix calculation

Now, we derive the transition probability matrix $P(t)$ for any t. We can calculate $p_{i,j}(t+\Delta t)$ as the probability of nucleotide i changing within time t to nucleotide k, and nucleotide k changing in the infinitesimally small time interval Δt from k to j, summed over all k. In a formula, this is

$$p_{i,j}(t+\Delta t) = \sum_{k=1}^{4} p_{i,k}(t) p_{k,j}(\Delta t). \tag{5.15}$$

In matrix notation, this is

$$P(t+\Delta t) = P(t)P(\Delta t). \tag{5.16}$$

Note that the summation over k illustrates that we take into account all intermediate substitutions when calculating the transition probability from i to j.

Since $P(t)P(\Delta t) = P(t) + P(t)Q\Delta t + \mathcal{O}(\Delta t^2)$, we obtain the following *difference equation* (which is the discrete-time analogue of a differential equation where time is continuous):

$$\frac{P(t+\Delta t) - P(t)}{\Delta t} = P(t)Q + \mathcal{O}(\Delta t). \tag{5.17}$$

If we take the limit $\Delta t \to 0$, we obtain the following differential equation:

$$\lim_{\Delta t \to 0} \frac{P(t+\Delta t) - P(t)}{\Delta t} = \frac{\mathrm{d}P}{\mathrm{d}t}(t) = P(t)Q. \tag{5.18}$$

Note that such a *differential equation* is also called a *master equation*: a master equation describes the time evolution of the probability of a system to occupy each one of a discrete set of states, with regard to a continuous time variable t.

Now, we need to solve this differential equation to find $P(t)$. By definition (Box 16 on page 88),

$$e^{Qt} = \sum_{i=0}^{\infty} \frac{(Qt)^i}{i!}. \tag{5.19}$$

Thus,

$$\frac{\mathrm{d}}{\mathrm{d}t} e^{Qt} = Q \sum_{i=1}^{\infty} i \frac{(Qt)^{i-1}}{i!} = Qe^{Qt}. \tag{5.20}$$

This means that

$$P(t) = e^{Qt} \tag{5.21}$$

is the solution of $\frac{\mathrm{d}}{\mathrm{d}t}P(t) = QP(t)$ with the initial value $P(0) = I$. Thus, the substitution rate matrix Q fully defines the transition probability matrix $P(t)$.

5.2.2.4 Evaluating the matrix exponential

We just showed that the substitution rate matrix Q fully defines the transition probability matrix $P(t)$ (Equation (5.21)). However, it is not clear how the matrix exponential $P(t) = e^{Qt}$ is evaluated in practice. Importantly, the exponential of a matrix is defined in terms of its Taylor expansion:

$$e^{Qt} = \sum_{i=0}^{\infty} \frac{(Qt)^i}{i!}. \tag{5.22}$$

We could evaluate this sum up to a very large i. However, this is numerically unstable and very slow since it requires many matrix multiplications. If Q is diagonalisable (see Box 19 on page 94), we can instead employ a matrix diagonalisation algorithm to find U and D such that D is a diagonal matrix containing the eigenvalues of Q and

$$Q = UDU^{-1}. \tag{5.23}$$

Since $(UDU^{-1})^n = UD^nU^{-1}$, substituting this into the Taylor series yields

$$\begin{aligned} e^{Qt} &= \sum_{n=0}^{\infty} \frac{UD^nt^nU^{-1}}{n!} \\ &= U\left(\sum_{n=0}^{\infty} \frac{D^nt^n}{n!}\right)U^{-1}. \end{aligned} \tag{5.24}$$

Furthermore, since Dt is diagonal, e^{Dt} is also diagonal with $(e^{Dt})_{i,i} = e^{D_{i,i}t}$. The exponentiated rate matrix is then simply

$$e^{Qt} = Ue^{Dt}U^{-1} = P(t). \tag{5.25}$$

Thus, in summary, given Q is diagonalisable, we can evaluate e^{Qt} by first determining U and D (by determining eigenvectors and eigenvalues of Q) and then taking the exponential of scalars (the diagonal elements of Dt) and two matrix multiplications ($Ue^{Dt}U^{-1}$). If Q is not diagonalisable, matrix exponentiation is difficult (Moler and Van Loan 1978; Moler and Van Loan 2003), and models with such rate matrices are rarely employed.

Box 19: Diagonalisable matrix

Before we define a diagonalisable matrix, we need some further terminology from linear algebra. A square matrix M of dimensions $n \times n$ is called a *diagonal matrix* if it has the form:

$$M = \begin{pmatrix} m_1 & 0 & \cdots & 0 \\ 0 & m_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & m_n \end{pmatrix}. \tag{B19.1}$$

If all diagonal entries of a diagonal matrix are 1, the matrix is called an *identity matrix* and is denoted I. A square matrix M of dimension $n \times n$ is called *invertible* if there exist an $n \times n$ dimensional matrix N such that

$$MN = I. \tag{B19.2}$$

N is often also noted as M^{-1}.

We can now define a diagonalisable matrix: an $n \times n$ dimensional matrix M is called *diagonalisable* if there exist a diagonal matrix D and an invertible matrix N both of dimensions $n \times n$ such that

$$M = NDN^{-1}. \tag{B19.3}$$

The entries on the diagonal of D are the eigenvalues of M.

A matrix M is called *symmetric* if $m_{i,j} = m_{j,i}$ for all i, j. The transpose M^T of a matrix M is obtained by flipping the entries of M over the diagonal, and thus, a matrix M is symmetric if and only if $M^T = M$.

Any symmetric matrix M is diagonalisable. If M only contains real values, then D and N only contain real values. We refer the reader to a textbook on linear algebra for the proofs.

5.2.3 Stochastic process of sequence evolution

We modelled the substitution process as an event happening in time step Δt with probability $r\Delta t$. We may make the simplifying assumption that the probability of more than one event is 0. This is a stochastic process in discrete time, with an experiment in each time step Δt having outcome “substitution” (probability $r\Delta t$) or “no substitution” (probability $1 - r\Delta t$). The process is called Bernoulli process (Box 20 on page 95), with the time to the first success (substitution) distributed geometrically (Box 15 on page 86), and the number of successes in L time steps distributed binomially (Box 3 on page 25).

When taking the continuous-time limit by letting Δt go to 0, we obtain the Poisson process (Box 21 on page 96) with time to the first event distributed exponentially (Section 5.2.2.1) and the number of events within a time interval of fixed length distributed according to the Poisson distribution (Box 21 on page 96).

The exponential distribution can directly be obtained as a limit of the geometric distribution (Box 18 on page 90), and the Poisson distribution as a limit of the binomial distribution

Box 20: Bernoulli process

A *Bernoulli process* is a sequence of independent Bernoulli trials with success probability p. In such a process, the probability of the first success occurring on the mth trial is given by a geometric distribution (Box 15 on page 86):

$$P(m|p) = p(1-p)^{m-1}. \tag{B20.1}$$

The probability that the Bernoulli process produces k successes after M total trials is given by the binomial distribution (Box 3 on page 25):

$$P(k|M,p) = \binom{M}{k} p^k (1-p)^{M-k}. \tag{B20.2}$$

(Box 22 on page 97), for an overview, see Box 23 on page 98.

5.2.4 Markov chain model of sequence evolution

The model of sequence evolution defined by the rate matrix Q that gives rise to transition probabilities as specified in Equation (5.21), is a *Markov chain*. Markov chains are a very common and useful stochastic process. These models have the convenient property of being *memoryless*, meaning that the probability of going from one state to another only depends on the current state and not on any previous state. Box 24 on page 98 explains the mathematical theory of Markov chains in a nutshell.

We can now observe that our sequence evolution model is a time-homogeneous Markov chain as it fulfils the following conditions:

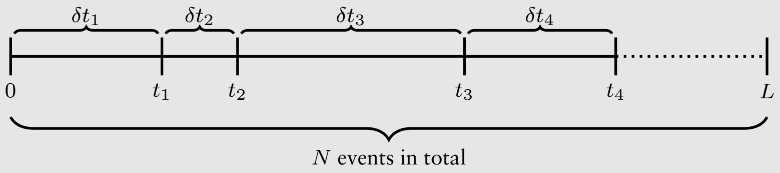
1. the state space we use in substitution models is finite; for example, for the nucleotide models, the state space is defined by $\mathcal{S} = \{\mathtt{T}, \mathtt{C}, \mathtt{A}, \mathtt{G}\}$;
2. the process is memoryless as the probability of substitution only depends on the current nucleotide i (via $q_{i,j}$), but not on the substitution history;
3. the process is time-homogenous as the rate matrix is identical for all times t.

Since the evolutionary models assume that each position in our MSA evolves independently from others, every position in our MSA is a separate Markov chain. Such a chain is illustrated in Figure 5.1.

Stationary distribution The *stationary distribution* π of a Markov chain is the distribution on the state space $\mathcal{S}$ which remains unchanged if the Markov chain acts on it further, meaning $\pi = \pi P(t)$. In this context, we define an *irreducible* and an *aperiodic* Markov chain.

Box 21: Poisson process

A *Poisson process* produces a sequence of events at a fixed rate r:



The waiting times between successive events in a Poisson process with rate r (the δt_i intervals in the above diagram) are exponentially distributed:

$$f_{\delta t_i}(x|r) = re^{-rx}. \tag{B21.1}$$

The probability that the process will generate N events in an interval of length L (as shown in the above diagram) is given by the *Poisson distribution*:

$$P(N = n|rL) = e^{-rL}\frac{(rL)^n}{n!}. \tag{B21.2}$$

The Poisson distribution has a mean and variance both equal to rL.

The Poisson process can be regarded as a continuous time limit of the Bernoulli process (Box 23 on page 98).

Combining the event times of two independent Poisson processes with rates r_A and r_B produces another Poisson process with rate $r_A + r_B$. This can be visualised by superimposing the events of the two processes on a single time axis, then noting that the time interval between events is given by the minimum of two exponentially distributed random variables with rates r_A and r_B. As explained in Box 17 on page 89, this random variable is also exponentially distributed with rate parameter $r_A + r_B$.

Similarly, a single Poisson process with rate r in which events are individually labelled A with probability q and B with probability $1 - q$ can be regarded as the union of two independent Poisson processes with rates rq and $r(1 - q)$, respectively. Decomposing Poisson processes in this fashion is known as thinning.

Irreducible means that it is possible for the system to go from one state to any other given enough steps (there are no states that cannot be reached). Mathematically this is guaranteed if for any time step $t > 0$, we have $p_{i,j}(t) > 0$ for $i \neq j$ and $i, j \in \{\mathsf{T}, \mathsf{C}, \mathsf{A}, \mathsf{G}\}$.

Aperiodic means that for any time step $t > 0$, we have $p_{i,i}(t) > 0$ for $i \in \{\mathsf{T}, \mathsf{C}, \mathsf{A}, \mathsf{G}\}$.

An irreducible and aperiodic Markov chain with transition probability matrix $P(t)$ has a unique stationary distribution; furthermore, $\lim_{t\to\infty} P(t)$ converges to a matrix where each row is the stationary distribution. This fact is known as the “Fundamental theorem of Markov Chains”.

The introduced substitution models are irreducible and aperiodic and thus have a stationary

Box 22: Connection between Poisson and binomial distribution

Here, we demonstrate that a limit of the binomial distribution is the Poisson distribution.

Consider the binomial distribution introduced in Box 3 on page 25:

$$P(k|M,p) = \binom{M}{k} p^k (1-p)^{M-k}. \tag{B22.1}$$

We define $r = {}^{pM}/_{L}$. For fixed $L > 0$, we evaluate the limit,

$$\begin{aligned}\lim_{M\to\infty} P(k|M, p = \frac{rL}{M}) &= \lim_{M\to\infty} \binom{M}{k} \left(\frac{rL}{M}\right)^k \left(1 - \frac{rL}{M}\right)^{M-k} \\ &= \frac{(rL)^k}{k!} \lim_{M\to\infty} \frac{M!}{(M-k)!M^k} \cdot \left(1 - \frac{rL}{M}\right)^{M-k}. \end{aligned} \tag{B22.2}$$

Since the limit of a product equals the product of the limits (provided the limits exist), we rewrite the right-hand side as

$$\frac{(rL)^k}{k!} \mathcal{L}_1 \cdot \mathcal{L}_2, \tag{B22.3}$$

where

$$\begin{aligned}\mathcal{L}_1 &= \lim_{M\to\infty} \frac{M!}{(M-k)!} \frac{1}{M^k} \\ &= \lim_{M\to\infty} \frac{M(M-1)\dots(M-k+1)}{M^k} = 1, \end{aligned} \tag{B22.4}$$

and

$$\begin{aligned}\mathcal{L}_2 &= \lim_{M\to\infty} \left(1 - \frac{rL}{M}\right)^{M-k} \\ &= \lim_{M\to\infty} \left(1 - \frac{rL}{M}\right)^{M} \lim_{M\to\infty} \left(1 - \frac{rL}{M}\right)^{-k} \\ &= \exp\left(\lim_{M\to\infty} M \log\left(1 - \frac{rL}{M}\right)\right) \\ &= \exp\left(\lim_{x\to 0} \frac{\log(1 - rLx)}{x}\right), \end{aligned} \tag{B22.5}$$

where $x = {}^{1}/_{M}$.

Since both the numerator and denominator of $\frac{\log(1-rLx)}{x}$ approach zero as x goes to zero, we can apply L'Hôpital's rule to find $\mathcal{L}_2 = e^{-rL}$ and thus that,

$$\lim_{M\to\infty} P(k|M, p = \frac{rL}{M}) = e^{-rL} \frac{(rL)^k}{k!}, \tag{B22.6}$$

which is the Poisson distribution with rate r.

In the context of Chapter 5, the number of substitutions within L discrete time steps of length Δt, where more than two substitutions within a time step have probability 0, follows the binomial distribution. When letting Δt go to 0 (making time continuous), the number of substitutions in an interval of a particular length ($L\Delta t$) follows the Poisson distribution. With the proof in this box, we showed that the continuous-time model is a limit of the discrete-time model.

Box 23: Relationships between different distributions

In previous boxes, we considered a sequence of Bernoulli trials in discrete time, meaning a Bernoulli process, and calculated the time to first success and the number of successes among a given number of trials. We further took the continuous-time limit of the sequence of trials, resulting in a Poisson process. The corresponding distributions are shown in the following table.

Bernoulli trials	Starting distribution	Limiting distribution	Limit
Time to first success	$m \sim$ Geometric(p)	$t \sim$ Exponential(r)	Constant $tr = mp$ where $m \to \infty$ and $p \to 0$.
Number of successes	$k \sim$ Binomial(p, M)	$k \sim$ Poisson(rL)	Constant $rL = Mp$ where $M \to \infty$ and $p \to 0$.

Box 24: Markov chain

A stochastic process is a series of random experiments performed through time. Time can be measured in discrete time steps or can be continuous. Consider a stochastic process that describes transitions ("jumps") between different states of the state space $\mathcal{S}$, where $\mathcal{S}$ is a finite or countable set. Such a stochastic process is a *Markov chain* if the probability of jumping from one state to another only depends on the current state and is independent of the history of past states. This is called the *Markov property*.

The mathematically rigorous definition of a Markov chain is as follows. Given state space $\mathcal{S}$ and a stochastic process $(X_t)_{t\in\mathcal{T}}$, where $\mathcal{T}$ is a discrete or continuous set of times, the process is called a Markov chain, if

$$P(X_{t_{n+1}} = x_{t_{n+1}} | X_{t_n} = x_{t_n}, X_{t_{n-1}} = x_{t_{n-1}}, \ldots) = P(X_{t_{n+1}} = x_{t_{n+1}} | X_{t_n} = x_{t_n}) \tag{B24.1}$$

holds for all $t_1 < t_2 < \ldots < t_n < t_{n+1}$ and $x_{t_1}, \ldots, x_{t_{n+1}} \in \mathcal{S}$.

This condition means that the state dynamics are *memoryless*: the state at time t_{n+1} only depends on the state at time t_n, but not on state $t_1, \ldots, t_{n-1}$, meaning the process has no memory of $t_1, \ldots, t_{n-1}$. In other words, the state we are in at the moment is the only one that matters for the next step of the Markov chain. We note that both the Bernoulli process and the Poisson process are Markov chains and, thus, are memoryless.

The Markov chain is called *time-homogeneous* if the probabilities on the state space do not change over time, that is, if $P(X_{t+h} = x_1 | X_t = x_0)$ is the same for all $t > 0$.

The process is called *stationary* if $(X_{t_1}, X_{t_2}, \ldots, X_{t_n})$ and $(X_{t_1+\tau}, X_{t_2+\tau}, \ldots, X_{t_n+\tau})$ have the same distribution for all $t_1, t_2, \ldots, t_n, \tau \in \mathcal{T}$.

A Markov chain with a finite number of states in state space $\mathcal{S}$ can be uniquely defined by the transition probability matrix P, and the P matrix is directly defined by Q, the rate matrix. Consult Kelly (1979) and Ross (1996) for further information on the theory of Markov chains.

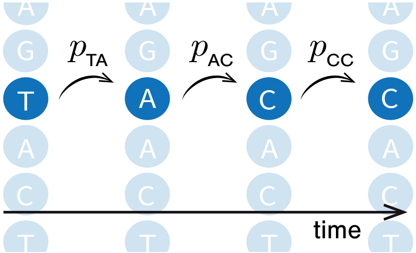


Figure 5.1: The changes of nucleotides, codons, and amino acids, resulting in molecular evolution, is modelled by Markov chains. Here, we provide an example using a nucleotide sequence. The vertical string of nucleotides is a sequence at a particular time. At each time point, the sequence is in a certain state and transitions to another state with probabilities defined by the transition probability matrix. Here, the sequence at the highlighted site started in state T and then changed to A with probability p_{TA}. It then changed to C and stayed there for the rest of the Markov chain process.

distribution. In our context, the stationary distribution is interpreted as follows. Given that we start with some arbitrary sequence, if we allow the sequence to evolve long enough under a substitution model with a rate matrix Q, the proportion of nucleotides in the evolved sequence will converge to the stationary distribution. Moreover, if the sequence evolves past that point, these proportions will not change further. We denote the probabilities of the four nucleotides of the stationary distribution with $\pi_{\mathsf{T}}, \pi_{\mathsf{C}}, \pi_{\mathsf{A}}, \pi_{\mathsf{G}}$. These probabilities are also called *equilibrium or stationary frequencies*, as in expectation the evolved sequence has a fraction/frequency of π_{T} Ts, π_{C} Cs, π_{A} As, and π_{G} Gs. This stationary distribution, and thus the equilibrium frequencies, is the same regardless of the starting sequence.

5.3 Common nucleotide substitution models

This section will discuss the most commonly used nucleotide substitution models and provide the rate matrices Q defining these models. Recall that we previously showed that the transition probability matrix can, in general, be obtained as $P(t) = e^{Qt}$ (Equation (5.21)). In what follows, we explicitly provide the matrix exponential for some rate matrices. Our notation and derivations are very similar to Yang (2014).

5.3.1 JC69 model

Substitution rate matrix under JC69 In the Jukes-Cantor (*JC69*) model, shown in Figure 5.2, all the substitutions occur at the same rate λ (Jukes and Cantor 1969). The substitu-

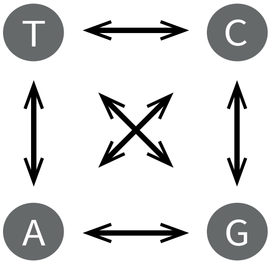


Figure 5.2: Schematic representation of the substitution rates of the JC69 model. The widths of the arrows represent the rates at which different substitutions happen. In the JC69 model, all rates are equal; thus, all arrows have the same width.

tion rate matrix Q_{JC69} is

$$Q_{\text{JC69}} = \begin{array}{c} \\ \mathsf{T} \\ \mathsf{C} \\ \mathsf{A} \\ \mathsf{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathsf{T} & \mathsf{C} & \mathsf{A} & \mathsf{G} \end{array} \\ \begin{pmatrix} -3\lambda & \lambda & \lambda & \lambda \\ \lambda & -3\lambda & \lambda & \lambda \\ \lambda & \lambda & -3\lambda & \lambda \\ \lambda & \lambda & \lambda & -3\lambda \end{pmatrix} \end{array}. \tag{5.26}$$

Transition probability matrix under JC69 In Section 5.2.2.3, we showed that we could calculate the transition probability matrix by diagonalising the substitution rate matrix Q (using the notation $Qt = UDtU^{-1}$). We obtain $P(t) = U\,\text{diag}\,(e^{\epsilon_1 t}, e^{\epsilon_2 t}, e^{\epsilon_3 t}, e^{\epsilon_4 t})\,U^{-1}$, where $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4$ are the eigenvalues of Q_{JC69}. The expression for $P(t)$ can be further rewritten as:

$$P(t) = \begin{pmatrix} p_0(t) & p_1(t) & p_1(t) & p_1(t) \\ p_1(t) & p_0(t) & p_1(t) & p_1(t) \\ p_1(t) & p_1(t) & p_0(t) & p_1(t) \\ p_1(t) & p_1(t) & p_1(t) & p_0(t) \end{pmatrix}, \tag{5.27}$$

where $p_0(t) = \frac{1}{4} + \frac{3}{4}e^{-4\lambda t}$ and $p_1(t) = \frac{1}{4} - \frac{1}{4}e^{-4\lambda t}$. We encourage the interested reader to derive these expressions by hand following the diagonalisation scheme described above. Note that these expressions only depend on one variable, namely λt, rather than the two variables λ and t. This makes intuitive sense: if we half the time but double the speed (the rate), we obtain the same outcome (see also Section 5.3.5).

Stationary distribution under JC69 For JC69, the stationary distribution is $\Pi = \{\pi_{\mathsf{T}}, \pi_{\mathsf{C}}, \pi_{\mathsf{A}}, \pi_{\mathsf{G}}\} = \{0.25, 0.25, 0.25, 0.25\}$. In general, as explained above, this is $\lim_{t\to\infty} P(t)$, and for the JC69 model leads to $\lim_{t\to\infty} p_0(t) = \lim_{t\to\infty} p_1(t) = 0.25$.

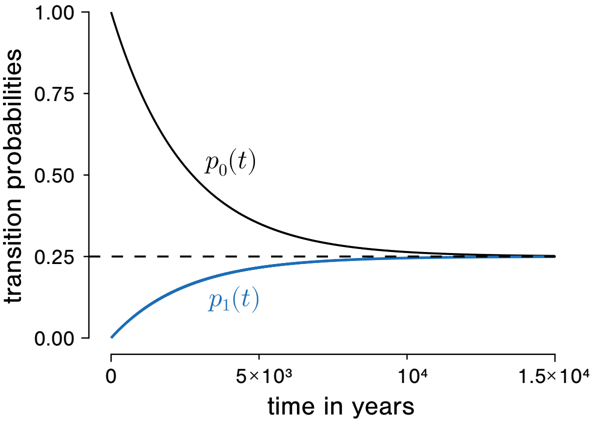


Figure 5.3: Changes in the transition probabilities of the JC69 model with $\lambda = 10^{-4}$ substitutions/site/year through time. The longer the process is running, the closer the probabilities are to 0.25. This is because the system approaches the stationary distribution after a long enough time.

Example Let us consider a virus genome evolving according to the JC69 model and assume that the substitution rate is $\lambda = 10^{-4}$ substitutions/site/year. The probability that we start with T and end up in C after $t = 10$ years is $p_{\mathsf{TC}}(10) = p_1(10) = 9.98 \times 10^{-4}$. The probability that T does not change in the same time step is $p_{\mathsf{TT}}(10) = 0.997006$. In this example, $\lambda t = 10^{-3}$, meaning that the approximation for small time steps (10^{-3}) and the exact transition probabilities (9.98×10^{-4}) are almost the same. However, if we are interested in $t = 10^4$, then $p_{\mathsf{TC}}(10^4) = 0.245$ while $\lambda t = 1$, meaning that using λt as an approximation of $p_{\mathsf{TC}}(t)$ is too imprecise, and we need to use the actual transition probability matrix.

The change of $p_0(t)$ and $p_1(t)$ with time follows the path shown in Figure 5.3. Starting from any nucleotide, if we let the sequence evolve for a long enough time, the probability of each of the four nucleotides at that site will be 0.25 (marked with a dashed line). Note that if we set λ to a value different from 10^{-4}, say, different by a factor f, then Figure 5.3 looks identical up to scaling the time axis by $^1/_f$.

The transition probability matrix $P(t)$ at different time steps is shown in Figure 5.4. The chain reached stationarity at the last observed time point $t = 2 \times 10^4$ years. Each site is equally likely to be in one of the four nucleotides, and the final sequence is, in expectation, composed of the four nucleotides in equal frequencies. Thus, if the analysed sequences are too divergent, meaning that the nucleotide content in each has reached the stationary distribution (saturation) due to long evolutionary time scales, it is impossible to calculate the relatedness between these sequences. They would all appear completely unrelated to each other.

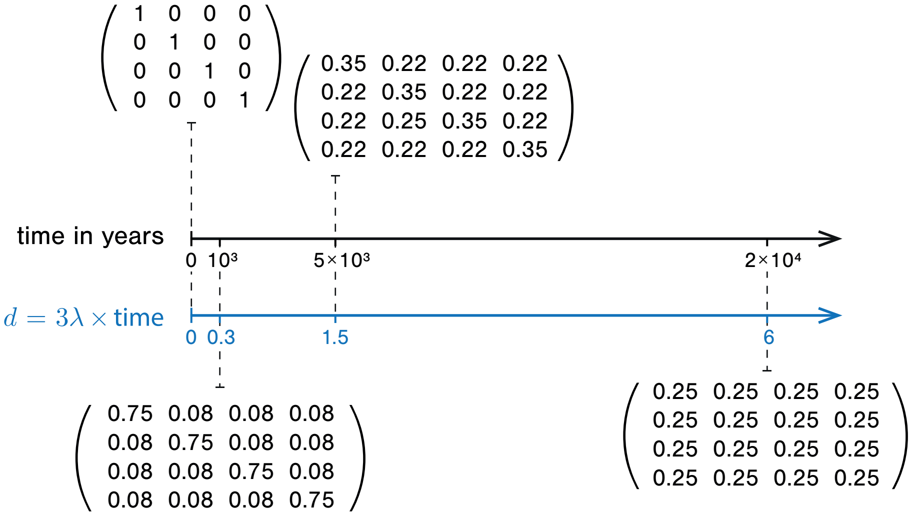


Figure 5.4: The changes in the transition probability matrix of the JC69 model with increasing time for the example substitution rate of $\lambda = 10^{-4}$ substitutions/site/year. At the beginning of the process $t = 0$, the sequence has not changed, and all the transition probabilities away from the original state (the off-diagonal entries of the transition probability matrix) are zero. After 2×10^4 years, all transition probabilities are 0.25. The time axis in years is displayed in black, while the time axis in units of substitutions is displayed in blue (see Section 5.3.5).

5.3.2 K80 model

Substitution rate matrix under K80 By studying real biological samples, scientists found that not all substitutions occur at the same rate. Substitutions between nucleotides with similar chemical structures are more likely than between two different structures. Thymine (T) and cytosine (C) consist of only one so-called pyrimidine ring structure; thus, they are referred to as *pyrimidines*. Adenine (A) and guanine (G) are purine derivatives, which consist of two rings (a pyrimidine ring fused to an imidazole ring); thus, they are named *purines*. Substitutions between two pyrimidines or two purines are referred to as *transitions*. Substitutions between one pyrimidine and one purine are referred to as *transversions*.

All substitution rates in JC69 are equal; thus, this model does not account for the difference in rates for transitions and transversions. Kimura (1980) extended the JC69 substitution rate model by accounting for differences between transitions and transversions, meaning that the substitutions between two purines ($\mathtt{A} \leftrightarrow \mathtt{G}$) and between two pyrimidines ($\mathtt{C} \leftrightarrow \mathtt{T}$) happen more easily and more often than the transversions, the substitutions between purines and pyrimidines ($\mathtt{A} \leftrightarrow \mathtt{C}$, $\mathtt{A} \leftrightarrow \mathtt{T}$, $\mathtt{G} \leftrightarrow \mathtt{C}$, $\mathtt{G} \leftrightarrow \mathtt{T}$), see Figure 5.5. This model is called *K80*.

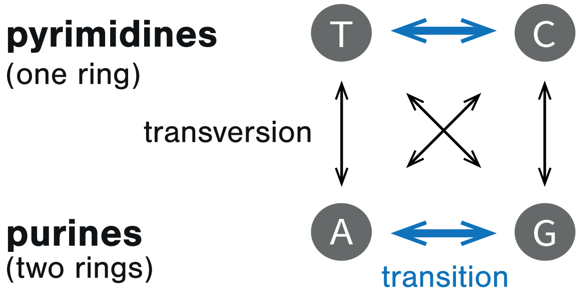


Figure 5.5: Schematic representation of the substitution rates of the K80 model. The widths of the arrows represent the rates at which different substitutions happen.

The substitution rate matrix Q_{K80} therefore contains two parameters, α for the transitions and β for the transversions:

$$Q_{\mathrm{K80}} = \begin{array}{c} \\ \mathsf{T} \\ \mathsf{C} \\ \mathsf{A} \\ \mathsf{G} \end{array} \begin{array}{c} \begin{array}{cccc} \quad\mathsf{T} & \qquad\qquad\mathsf{C} & \qquad\qquad\mathsf{A} & \qquad\qquad\mathsf{G} \end{array} \\ \begin{pmatrix} -(\alpha+2\beta) & \alpha & \beta & \beta \\ \alpha & -(\alpha+2\beta) & \beta & \beta \\ \beta & \beta & -(\alpha+2\beta) & \alpha \\ \beta & \beta & \alpha & -(\alpha+2\beta) \end{pmatrix} \end{array}. \tag{5.28}$$

Transition probability matrix under K80 Using the same diagonalisation procedure as for JC69 we can calculate the transition probability matrix $P(t) = e^{Qt}$:

$$P(t) = \begin{pmatrix} p_0(t) & p_1(t) & p_2(t) & p_2(t) \\ p_1(t) & p_0(t) & p_2(t) & p_2(t) \\ p_2(t) & p_2(t) & p_0(t) & p_1(t) \\ p_2(t) & p_2(t) & p_1(t) & p_0(t) \end{pmatrix}, \tag{5.29}$$

where

$$p_0(t) = \frac{1}{4} + \frac{1}{4}e^{-4\beta t} + \frac{1}{2}e^{-2(\alpha+\beta)t}, \tag{5.30}$$

$$p_1(t) = \frac{1}{4} + \frac{1}{4}e^{-4\beta t} - \frac{1}{2}e^{-2(\alpha+\beta)t}, \tag{5.31}$$

$$p_2(t) = \frac{1}{4} - \frac{1}{4}e^{-4\beta t}. \tag{5.32}$$

As an exercise, we advise the interested reader to derive these probabilities using pen and paper.

Analogously to the JC69 model, these equations only depend on two (α, β) instead of three

(α, β, t) variables. The K80 model is thus very often parameterised in terms of the distance between two sequences separated by time t, $d = (\alpha + 2\beta)t$ (Section 5.3.5), and the ratio between the transition and transversion rates $\kappa = {}^{\alpha}/_{\beta}$. With these definitions, the transition probabilities transform into:

$$p_0(t) = \frac{1}{4} + \frac{1}{4}e^{-4d/(\kappa+2)} + \frac{1}{2}e^{-2d(\kappa+1)/(\kappa+2)}, \tag{5.33}$$

$$p_1(t) = \frac{1}{4} + \frac{1}{4}e^{-4d/(\kappa+2)} - \frac{1}{2}e^{-2d(\kappa+1)/(\kappa+2)}, \tag{5.34}$$

$$p_2(t) = \frac{1}{4} - \frac{1}{4}e^{-4d/(\kappa+2)}. \tag{5.35}$$

Stationary distribution under K80. Again, by letting t go to infinity in the expressions $p_i(t), i \in \{0, 1, 2\}$, we find that all nucleotides have stationary frequencies of 0.25.

As a short side note, we want to remind the reader of the order we chose for the nucleotides in the nucleotide substitution rate matrix. By using the order T, C, A, G, transitions cluster together in the substitution rate matrix, which would not be the case for an alphabetical arrangement A, C, G, T.

5.3.3 F81 model

Substitution rate matrix under F81 *F81* is another extension of the JC69 model in which the equilibrium frequencies can deviate from 0.25 (Felsenstein 1981). Recall that equilibrium frequencies describe the expected fraction of T, C, A, and G in the sequence after the stationary distribution, an evolutionary equilibrium, is reached. Under F81, the equilibrium frequencies can take any values $\pi_N \in [0, 1], N \in \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$ where $\pi_\mathsf{A} + \pi_\mathsf{C} + \pi_\mathsf{G} + \pi_\mathsf{T} = 1$.

The substitution rate matrix is defined as:

$$Q_{\mathrm{F81}} = \begin{array}{c} \\ \mathsf{T} \\ \mathsf{C} \\ \mathsf{A} \\ \mathsf{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathsf{T} & \mathsf{C} & \mathsf{A} & \mathsf{G} \end{array} \\ \begin{pmatrix} -(\pi_\mathsf{C} + \pi_\mathsf{A} + \pi_\mathsf{G}) & \pi_\mathsf{C} & \pi_\mathsf{A} & \pi_\mathsf{G} \\ \pi_\mathsf{T} & -(\pi_\mathsf{T} + \pi_\mathsf{A} + \pi_\mathsf{G}) & \pi_\mathsf{A} & \pi_\mathsf{G} \\ \pi_\mathsf{T} & \pi_\mathsf{C} & -(\pi_\mathsf{T} + \pi_\mathsf{C} + \pi_\mathsf{G}) & \pi_\mathsf{G} \\ \pi_\mathsf{T} & \pi_\mathsf{C} & \pi_\mathsf{A} & -(\pi_\mathsf{T} + \pi_\mathsf{C} + \pi_\mathsf{A}) \end{pmatrix} \end{array}. \tag{5.36}$$

Transition probability matrix under F81 F81 is a special case of the TN93 model, which we will discuss in Section 5.3.4. The transition probability matrix for F81 is stated in Yang (2014, Equation 1.20). Letting t go to infinity in the entries of the transition probability matrix confirms that the equilibrium frequencies are indeed $\pi_\mathsf{A}, \pi_\mathsf{C}, \pi_\mathsf{G}, \pi_\mathsf{T}$.

5.3.4 More general nucleotide substitution models

In the following, we present more general models and focus on their substitution rate matrices. For further properties of the models, we refer the interested reader to Yang (2014) and Felsenstein (2003).

Hasegawa, Yano, and Kishino extended the K80 and F81 models to account both for transitions and transversions and arbitrary equilibrium frequencies of the nucleotides (Hasegawa, Yano and Kishino 1984). The model, normally referred to as *HKY*, has the substitution rate matrix:

$$Q_{\mathrm{HKY}} = \begin{array}{c} \\ \mathsf{T} \\ \mathsf{C} \\ \mathsf{A} \\ \mathsf{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathsf{T} & \mathsf{C} & \mathsf{A} & \mathsf{G} \end{array} \\ \begin{pmatrix} -(\alpha\pi_{\mathsf{C}} + \beta\pi_{\mathsf{A}} + \beta\pi_{\mathsf{G}}) & \alpha\pi_{\mathsf{C}} & \beta\pi_{\mathsf{A}} & \beta\pi_{\mathsf{G}} \\ \alpha\pi_{\mathsf{T}} & -(\alpha\pi_{\mathsf{T}} + \beta\pi_{\mathsf{A}} + \beta\pi_{\mathsf{G}}) & \beta\pi_{\mathsf{A}} & \beta\pi_{\mathsf{G}} \\ \beta\pi_{\mathsf{T}} & \beta\pi_{\mathsf{C}} & -(\beta\pi_{\mathsf{T}} + \beta\pi_{\mathsf{C}} + \alpha\pi_{\mathsf{G}}) & \alpha\pi_{\mathsf{G}} \\ \beta\pi_{\mathsf{T}} & \beta\pi_{\mathsf{C}} & \alpha\pi_{\mathsf{A}} & -(\beta\pi_{\mathsf{T}} + \beta\pi_{\mathsf{C}} + \alpha\pi_{\mathsf{A}}) \end{pmatrix} \end{array}. \tag{5.37}$$

Tamura and Nei (1993) introduced a yet more sophisticated model, called *TN93*, where the rates of T ↔ C transitions (α_1) can be different from those of A ↔ G (α_2). The substitution rate matrix under TN93 is:

$$Q_{\mathrm{TN93}} = \begin{array}{c} \\ \mathsf{T} \\ \mathsf{C} \\ \mathsf{A} \\ \mathsf{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathsf{T} & \mathsf{C} & \mathsf{A} & \mathsf{G} \end{array} \\ \begin{pmatrix} -(\alpha_1\pi_{\mathsf{C}} + \beta\pi_{\mathsf{A}} + \beta\pi_{\mathsf{G}}) & \alpha_1\pi_{\mathsf{C}} & \beta\pi_{\mathsf{A}} & \beta\pi_{\mathsf{G}} \\ \alpha_1\pi_{\mathsf{T}} & -(\alpha_1\pi_{\mathsf{T}} + \beta\pi_{\mathsf{A}} + \beta\pi_{\mathsf{G}}) & \beta\pi_{\mathsf{A}} & \beta\pi_{\mathsf{G}} \\ \beta\pi_{\mathsf{T}} & \beta\pi_{\mathsf{C}} & -(\beta\pi_{\mathsf{T}} + \beta\pi_{\mathsf{C}} + \alpha_2\pi_{)} & \alpha_2\pi_{\mathsf{G}} \\ \beta\pi_{\mathsf{T}} & \beta\pi_{\mathsf{C}} & \alpha_2\pi_{\mathsf{A}} & -(\beta\pi_{\mathsf{T}} + \beta\pi_{\mathsf{C}} + \alpha_2\pi_{\mathsf{A}}) \end{pmatrix} \end{array}. \tag{5.38}$$

Note that HKY is a special case of TN93 where $\alpha_1 = \alpha_2$.

The *general time-reversible model*, *GTR*, has become very popular (Tavaré 1986; Yang 1994; Zharkikh 1994):

$$Q_{\mathrm{GTR}} = \begin{array}{c} \\ \mathsf{T} \\ \mathsf{C} \\ \mathsf{A} \\ \mathsf{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathsf{T} & \mathsf{C} & \mathsf{A} & \mathsf{G} \end{array} \\ \begin{pmatrix} -(a\pi_{\mathsf{C}} + b\pi_{\mathsf{A}} + c\pi_{\mathsf{G}}) & a\pi_{\mathsf{C}} & b\pi_{\mathsf{A}} & c\pi_{\mathsf{G}} \\ a\pi_{\mathsf{T}} & -(a\pi_{\mathsf{T}} + d\pi_{\mathsf{A}} + e\pi_{\mathsf{G}}) & d\pi_{\mathsf{A}} & e\pi_{\mathsf{G}} \\ b\pi_{\mathsf{T}} & d\pi_{\mathsf{C}} & -(b\pi_{\mathsf{T}} + d\pi_{\mathsf{C}} + f\pi_{\mathsf{G}}) & f\pi_{\mathsf{G}} \\ c\pi_{\mathsf{T}} & e\pi_{\mathsf{C}} & f\pi_{\mathsf{A}} & -(c\pi_{\mathsf{T}} + e\pi_{\mathsf{C}} + f\pi_{\mathsf{A}}) \end{pmatrix} \end{array}. \tag{5.39}$$

This is the most general model that satisfies the time-reversibility constraint, which we discuss in detail in Section 5.3.6. All the previously discussed models are special cases of the GTR model.

The most general substitution model without any constraints on parameter values is called the *UNREST* (for "unrestricted") model (Yang 1994):

$$Q_{\mathrm{UNREST}} = \begin{array}{c} \\ \mathsf{T} \\ \mathsf{C} \\ \mathsf{A} \\ \mathsf{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathsf{T} & \mathsf{C} & \mathsf{A} & \mathsf{G} \end{array} \\ \begin{pmatrix} -(a+b+c) & a & b & c \\ d & -(d+e+f) & e & f \\ g & h & -(g+h+i) & i \\ j & k & l & -(j+k+l) \end{pmatrix} \end{array}. \tag{5.40}$$

Each substitution rate in this model can be different and is a separate parameter of the model. All previously discussed models are special cases of the UNREST model. The UNREST model is not time reversible in general, and mathematical derivations under it are generally very complicated.

Thus far, we have provided the rate matrices for these general substitution models. We now briefly discuss their transition probabilities. There are known analytical solutions for the transition probabilities of the HKY and TN93 substitution models, and interested readers are encouraged to refer to Yang (2014). There is no explicit analytical solution for the transition probabilities under the GTR model. However, Q_{GTR} is diagonalisable with real eigenvalues (see Section 5.3.6), enabling efficient and stable numerical diagonalisation. Obtaining the transition probabilities for the UNREST model is more difficult as its substitution rate matrix Q_{UNREST} is not generally diagonalisable.

Finally, the equilibrium probabilities for the GTR model (and for all models that are special cases of GTR) are $\pi_{\mathsf{T}}, \pi_{\mathsf{C}}, \pi_{\mathsf{A}}, \pi_{\mathsf{G}}$ (Yang 2014). For the UNREST model, we have the general property $\pi P_{\text{UNREST}}(t) = \pi$, which is equivalent to $\pi Q_{\text{UNREST}} = 0$ (Grimmett and Stirzaker 1992), meaning we can derive the equilibrium probabilities based on the UNREST substitution rate matrix by solving $\pi Q_{\text{UNREST}} = 0$.

5.3.5 Time scale: calendar time versus evolutionary time

Evolutionary processes can be measured in units of calendar time (days, years, and so on). However, one can also express time in terms of expected numbers of substitutions, called (expected) distance d, through a simple transformation:

$$d = \frac{\text{time}}{(\text{expected time until any substitution})}. \tag{5.41}$$

For the JC69 model, this would be $d = {}^{t}/{(3\lambda)^{-1}} = 3\lambda t$, the expected number of substitutions in elapsed time t. The advantage of using d is that it summarises two parameters (time and rate) into one quantity. We can estimate this quantity d, the expected number of substitutions that occurred when one sequence evolved into another one, by looking at the two sequences (details on that follow in Section 5.4). The distance d remains the same whether the sequences evolved at rate λ in time t or at rate 2λ in time ${}^{t}/{2}$.

To calculate the calendar time that passed while one sequence evolved into the other, we would need to know λ, which is typically unknown. In other words, obtaining separate estimates of calendar time and substitution rates is often impossible.

Note that the default output of many phylogenetic inference methods is actually the distance in units of substitutions (see the blue axis in Figure 5.4).

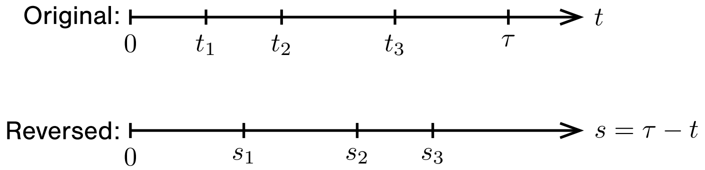


Figure 5.6: In defining a time-reversible process, consider the times t_1, t_2 and t_3 marked on the "Original" axis above. If time reversibility holds, the joint probability of states at these times should be exactly equal to the joint probability for these same states but with time reversed. The time reversal is achieved by choosing a time τ on the original axis and defining a transformed time variable $s = \tau - t$. This means that $s = 0$ on the time-reversed axis corresponds to $t = \tau$ on the original axis.

5.3.6 Time-reversibility of the nucleotide substitution models

A stochastic process $(X_t)_{t \in \mathcal{T}}$ is called *time-reversible* if it shows the same statistical behaviour forwards and backwards in time (Kelly 1979). Intuitively, one can picture this property in the following way: imagine filming a time-reversible stochastic process through time. Then, regardless of whether the recording is played backwards or forwards, the two depictions of the process will be statistically indistinguishable.

Almost all commonly used substitution models are time-reversible. The primary reason for this is convenience. In particular, time-reversible models have mathematical properties that make it easy to compute their transition probability matrices (see below and Felsenstein (2003)). Additionally, the probability of sequences given a phylogenetic tree does not depend on the position of the root of the tree (see Section 6.3.3). This means that phylogenetic inference algorithms can avoid a lot of computation when searching for the optimal tree (Boussau and Gouy 2006). We will make use of these properties when calculating the probability of observing specific sequences given a tree and a substitution model (tree likelihood calculation; Sections 6.3.3.1 and 6.3.3.2).

Formally, a stochastic process $(X_t)_{t \in \mathcal{T}}$ is called *time-reversible* if $(X_{t_1}, X_{t_2}, \ldots, X_{t_n})$ and $(X_{\tau - t_n}, X_{\tau - t_{n-1}}, \ldots, X_{\tau - t_1})$ have the same distribution for all $t_1, t_2, \ldots, t_n, \tau \in \mathcal{T}$ (see Figure 5.6 for the intuition behind this transformation). We call this the time-reversibility condition.

Lemma 5.3.1. *For a stationary Markov chain (see Box 24 on page 98) on state space $\mathcal{S}$ with transition probabilities $p_{i,j}$, rates $q_{i,j}$, and stationary probabilities π_i, $i, j \in \mathcal{S}$, the time-reversibility condition is equivalent to the following condition:*

$$\pi_i p_{i,j}(t) = \pi_j p_{j,i}(t), \tag{5.42}$$

and also to:

$$\pi_i q_{i,j} = \pi_j q_{j,i}. \tag{5.43}$$

A comprehensive proof of these equivalences can be found in Kelly (1979, Theorems 1.2 and 1.3).

Conditions shown in Equations (5.42) and (5.43) are also called *detailed balance conditions*. Intuitively, Equation (5.42) can be interpreted such that the probability flux from state i to j must equal the probability flux from state j to i.

Lemma 5.3.2. *Consider a Markov chain on state space $\mathcal{S}$ with transition probabilities $p_{i,j}$. Suppose a distribution π fulfils the detailed balance condition in Equation* (5.42). *Then, the given Markov chain has a stationary distribution, and this distribution is π.*

Proof. We have

$$\sum_i \pi_i p_{i,j} = \sum_i \pi_j p_{j,i} = \pi_j \sum_i p_{j,i} = \pi_j. \tag{5.44}$$

Thus, in matrix notation, $\pi P = \pi$, and by definition, π is a stationary distribution. □

In practice, we can determine the time-reversibility of a process easily from the rate matrix Q, as we show in Theorem 5.3.3.

Theorem 5.3.3. *A stationary Markov chain with rate matrix Q is time-reversible if and only if the rate matrix can be decomposed into a symmetric matrix $S = (s_{i,j})_{i,j \in \{1,2,\ldots,n\}}$ and a diagonal matrix Π. The equilibrium frequencies are on the diagonals of Π:*

$$Q = \begin{pmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,n} \\ s_{1,2} & s_{2,2} & \cdots & s_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{1,n} & s_{2,n} & \cdots & s_{n,n} \end{pmatrix} \cdot \begin{pmatrix} \pi_1 & 0 & \cdots & 0 \\ 0 & \pi_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \pi_n \end{pmatrix}. \tag{5.45}$$

Proof. To prove this statement, we will prove that Equations (5.43) and (5.45) are equivalent. We do so by showing that if Equation (5.43) holds, Equation (5.45) can be derived, and the other way around.

Deriving Equation (5.45) from Equation (5.43) Let us assume the Markov chain with rate matrix Q fulfils Equation (5.43), which implies that

$$q_{i,j} \overset{(5.43)}{=} \frac{\pi_j}{\pi_i} q_{j,i} = \pi_j \underbrace{\frac{1}{\pi_i} q_{j,i}}_{=s_{i,j}} = \pi_j s_{i,j} \tag{5.46}$$

holds for all $i, j \in \{1, 2, \ldots, n\}$. In this equation, we defined parameters $s_{i,j}$ such that

$$s_{i,j} = \frac{1}{\pi_i} q_{j,i}. \tag{5.47}$$

We can now rewrite the substitution rate matrix by replacing the entries according to Equation (5.46):

$$Q = \begin{pmatrix} q_{1,1} & q_{1,2} & \cdots & q_{1,n} \\ q_{2,1} & q_{2,2} & \cdots & q_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ q_{n,1} & q_{n,2} & \cdots & q_{n,n} \end{pmatrix} = \begin{pmatrix} \pi_1 s_{1,1} & \pi_2 s_{1,2} & \cdots & \pi_n s_{1,n} \\ \pi_1 s_{2,1} & \pi_2 s_{2,2} & \cdots & \pi_n s_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \pi_1 s_{n,1} & \pi_2 s_{n,2} & \cdots & \pi_n s_{n,n} \end{pmatrix}. \tag{5.48}$$

This can be decomposed into:

$$\begin{pmatrix} \pi_1 s_{1,1} & \pi_2 s_{1,2} & \cdots & \pi_n s_{1,n} \\ \pi_1 s_{2,1} & \pi_2 s_{2,2} & \cdots & \pi_n s_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \pi_1 s_{n,1} & \pi_2 s_{n,2} & \cdots & \pi_n s_{n,n} \end{pmatrix} = \begin{pmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,n} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n,1} & s_{n,2} & \cdots & s_{n,n} \end{pmatrix} \cdot \begin{pmatrix} \pi_1 & 0 & \cdots & 0 \\ 0 & \pi_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \pi_n \end{pmatrix}. \tag{5.49}$$

It remains to show that the S matrix is symmetric, $s_{i,j} = s_{j,i}$. This is trivial for $i = j$; thus, we assume now $i \neq j$. Then the symmetry follows from the definition of $s_{i,j}$ and the time-reversibility equation:

$$s_{i,j} \overset{(5.47)}{=} \frac{1}{\pi_i} q_{j,i} \overset{(5.43)}{=} \frac{1}{\pi_i} \frac{\pi_i}{\pi_j} q_{i,j} \overset{(5.47)}{=} s_{j,i}. \tag{5.50}$$

Deriving Equation (5.43) from Equation (5.45) Let us assume that the Markov chain with rate matrix Q fulfils Equation (5.45). For $i = j$, Equation (5.43) is always true. Thus, we now look at the case $i \neq j$. We have $s_{i,j} = s_{j,i}$, and thus we can show that Equation (5.43) holds:

$$\pi_i q_{i,j} = \pi_i s_{j,i} \pi_j = \pi_j s_{i,j} \pi_i = \pi_j q_{j,i}. \tag{5.51}$$

In summary, we proved that Equation (5.43) and Equation (5.45) are equivalent. □

As noted previously, the most general time-reversible substitution model is the GTR model. The time-reversibility can be seen from decomposing the Q_{GTR} matrix (Equation (5.39)) into

$$
\begin{aligned}
Q_{\mathrm{GTR}} &= \begin{pmatrix} -(a\pi_{\mathrm{C}} + b\pi_{\mathrm{A}} + c\pi_{\mathrm{G}}) & a\pi_{\mathrm{C}} & b\pi_{\mathrm{A}} & c\pi_{\mathrm{G}} \\ a\pi_{\mathrm{T}} & -(a\pi_{\mathrm{T}} + d\pi_{\mathrm{A}} + e\pi_{\mathrm{G}}) & d\pi_{\mathrm{A}} & e\pi_{\mathrm{G}} \\ b\pi_{\mathrm{T}} & d\pi_{\mathrm{C}} & -(b\pi_{\mathrm{T}} + d\pi_{\mathrm{C}} + f\pi_{\mathrm{G}}) & f\pi_{\mathrm{G}} \\ c\pi_{\mathrm{T}} & e\pi_{\mathrm{C}} & f\pi_{\mathrm{A}} & -(c\pi_{\mathrm{T}} + e\pi_{\mathrm{C}} + f\pi_{\mathrm{A}}) \end{pmatrix} \\
&= \begin{pmatrix} -\frac{a\pi_{\mathrm{C}}+b\pi_{\mathrm{A}}+c\pi_{\mathrm{G}}}{\pi_{\mathrm{T}}} & a & b & c \\ a & -\frac{a\pi_{\mathrm{T}}+d\pi_{\mathrm{A}}+e\pi_{\mathrm{G}}}{\pi_{\mathrm{C}}} & d & e \\ b & d & -\frac{b\pi_{\mathrm{T}}+d\pi_{\mathrm{C}}+f\pi_{\mathrm{G}}}{\pi_{\mathrm{A}}} & f \\ c & e & f & -\frac{c\pi_{\mathrm{T}}+e\pi_{\mathrm{C}}+f\pi_{\mathrm{A}}}{\pi_{\mathrm{G}}} \end{pmatrix} \times \begin{pmatrix} \pi_{\mathrm{T}} & 0 & 0 & 0 \\ 0 & \pi_{\mathrm{C}} & 0 & 0 \\ 0 & 0 & \pi_{\mathrm{A}} & 0 \\ 0 & 0 & 0 & \pi_{\mathrm{G}} \end{pmatrix}.
\end{aligned}
\tag{5.52}
$$

The other shown time-reversible substitution models, JC69, K80, F81, HKY, and TN93, are special cases of GTR.

As discussed in Section 5.2.2, it is straightforward to obtain $P(t)$ if Q is diagonalisable with real eigenvalues. Corrolary 5.3.4 thus shows that it is easy to calculate $P(t)$ for all time reversible models.

Corollary 5.3.4. *Let Q be the rate matrix of a stationary time-reversible Markov chain. Then Q is diagonalisable and has real eigenvalues.*

Proof. We can write:

$$Q \overset{(5.3.3)}{=} S\Pi = \Pi^{-1/2}\Pi^{1/2}S\Pi^{1/2}\Pi^{1/2}, \tag{5.53}$$

where S is a symmetric matrix and Π is a diagonal matrix.

From this, we can establish that $\Pi^{1/2}S\Pi^{1/2}$ is symmetric:

$$(\Pi^{1/2}S\Pi^{1/2})^T = (\Pi^{1/2})^T(S)^T(\Pi^{1/2})^T = \Pi^{1/2}S\Pi^{1/2}. \tag{5.54}$$

Since $\Pi^{1/2}S\Pi^{1/2}$ is symmetric, we can write $\Pi^{1/2}S\Pi^{1/2} = U^{-1}\Lambda U$ where Λ contains real eigenvalues (see Box 19 on page 94).

Then $Q = \Pi^{1/2}U^{-1}\Lambda U\Pi^{1/2} = (U\Pi^{1/2})^{-1}\Lambda(U\Pi^{1/2})$ where Λ contains the real eigenvalues of Q. □

Model	Parameters	Description
JC69	1	all substitutions have the same rate, all equilibrium frequencies are equal
K80	2	transitions and transversions have different rates, all equilibrium frequencies are equal
HKY	$2 + 3^*$	transitions and transversions have different rates, equilibrium frequencies can be different
TN93	$3 + 3^*$	as HKY, but different rates for different kinds of transitions
GTR	$6 + 3^*$	general, all rates and equilibrium frequencies can vary, but the model is still time-reversible
UNREST	12	most general, not time-reversible

Table 5.1: Overview of substitution rate models and their number of parameters. The numbers with * correspond to the number of free equilibrium frequency parameters that can either be co-estimated alongside the remaining parameters or be fixed based on independent data.

5.3.7 Site dependency in molecular substitution models

The overview of models mentioned in this chapter, the number of their parameters, and a short description are displayed in Table 5.1. All discussed models have in common that they assume that sites change independently from each other. It is debatable whether the assumption of independence between sites is justified, and ignoring dependence when it is present may lead to accuracy loss in analysis results (Nasrallah, Mathews and Huelsenbeck 2010). There are models that define transition probabilities for nucleotide triplets (codons) called *codon models* (Section 5.7), assuming that the individual nucleotides evolve together dictated by the properties of the corresponding amino acid (see also Section 5.6). However, only a little work has been done on any dependency beyond the nucleotides within codons, as accounting for dependence is computationally very hard: we would need to calculate the likelihood for each combination of states along a lineage for all sites simultaneously. For example, some work on accounting for particular site dependences can be found in Arndt and Hwa (2005) and Hoehn, Lunter and Pybus (2017).

5.4 Distance estimation for nucleotide sequences

One approach to reconstructing phylogenies is to calculate the distance d between all pairs of sequences and build a phylogeny by grouping sequences with small distances close together (see Section 6.3.1). In what follows, we will introduce methods to estimate distance d for pairs of sequences. That is, we will introduce estimators $\hat{d}$ for d (estimators are normally marked with $\hat{\cdot}$). We first explain why simply counting the number of differences between two sequences is not a good way to estimate d, and then derive estimators for d under the JC69 model. Table 5.2 states estimators for further models.

5.4.1 Simple pairwise distances

The *Hamming distance* and *p-distance* are the simplest measures of distances between two sequences of equal length. The Hamming distance is simply the number of *segregating sites*, the number of sites that vary between the two sequences. The p-distance is the Hamming distance divided by the total sequence length. For example, the Hamming distance between the following two sequences:

```
ACTAGCTG
AGTTGCTG
```

is 2, whereas the p-distance is $^2/_8 = 0.25$. In the example of *triose-phosphate isomerase* from Chapter 3, Figure 3.1, the Hamming distance between mosquito and rice sequences is 35, whereas the p-distance is 0.636.

Both the Hamming distance and the p-distance are very simplistic measures that ignore the fact that some substitutions are unobserved (or hidden) when considering the observed sequences. We will go through different scenarios of successive substitutions to illustrate cases where the Hamming and p-distance measures are biologically inadequate. We will use the tree shown in Figure 5.7 as an example. For demonstration, we assume that we know the true sequence at the internal node depicting the common ancestor of taxon 1 and taxon 2[1].

The daughter sequences (taxa 1 and 2) differ in two sites of the MSA. In the second site, the nucleotide difference could have resulted from a single substitution, as depicted in Figure 5.7 (orange). In this case, a C changed to a G on the branch leading to taxon 2. The Hamming and p-distance would correctly account for one substitution.

However, in our example, there have been several unobserved substitutions, which the Hamming or p-distance cannot account for. This is highlighted in blue and grey. First, we have *multiple substitutions* at a single position on a branch between the common ancestor and taxon 1. The T in position 4 (depicted in blue) of the common ancestor sequence was first replaced by a C and only later by an A, but all we can observe is the A in the final sequence. Here, the Hamming distance is 1, while two substitutions occurred in reality. Second, we may have *parallel substitutions*, meaning a site in both descendant sequences changes to the same nucleotide. In our example, the A in position 6 (grey) changed into C in parallel for taxa 1 and 2. Here, the Hamming distance is 0, while two substitutions occurred.

[1] A *taxon* (plural taxa) is a term used in the science of biological classification (which is referred to as *taxonomy*). It describes a biological unit defined in taxonomy and can refer to a group of one or more populations of an organism or organisms. Taxa are typically arranged in a hierarchy from kingdom to subspecies.

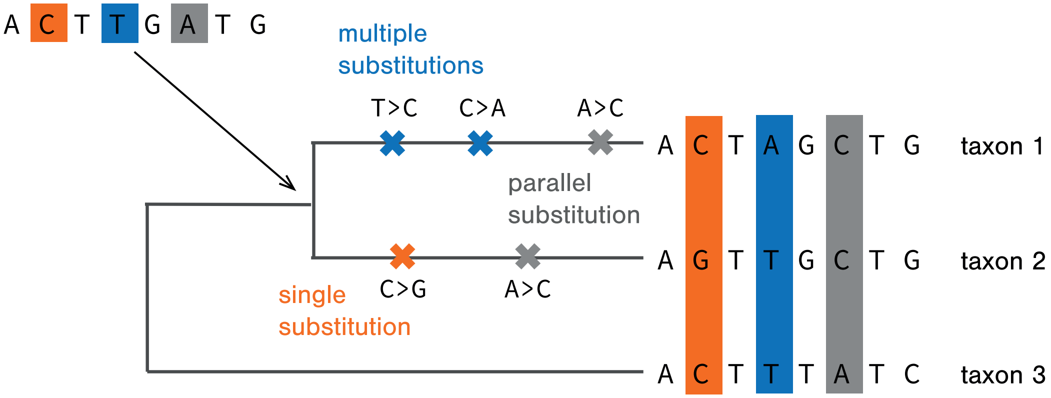


Figure 5.7: An illustration of different substitution scenarios on a phylogenetic tree with three tips. The arrow indicates the most recent common ancestor of taxa 1 and 2. **Orange:** A single substitution happened at site 2 on the branch leading to taxon 2. **Blue:** Two consecutive substitutions happened at site 4 on the branch leading to taxon 1. **Gray:** At site 6, taxa 1 and 2 each had one independent substitution from adenine (A) to cytosine (C).

Moreover, we also could not account for all substitutions using the Hamming or p-distance in case of *convergent substitution* (a certain position is different in two ancestral lineages, but both lineages have a substitution into the same nucleotide) or *back substitutions* (a site changes to a different nucleotide, and then changes back to the original starting nucleotide).

In summary, the Hamming and p-distance measure a minimal distance between sequences since they report the minimal changes required to evolve from one sequence into the other. Using such minimal distances may bias downstream results, and a distance measure that represents the actual evolutionary distance acknowledging potential hidden or unobserved substitutions is preferable. The mathematical models of sequence evolution introduced above provide such distance measures.

5.4.2 Pairwise distances for JC69 using the method of moments

Let us consider two sequences of length n each. One sequence was the starting sequence, having evolved for time t under the JC69 model, into the second sequence. We want to estimate their evolutionary distance, taking into account all possible hidden intermediate substitutions. We start with defining the probability of any substitution over time t. From the transition probability matrix, we obtain the total probability that a nucleotide changes:

$$p(t) = 3p_1(t) = \frac{3}{4} - \frac{3}{4}e^{-4\lambda t}. \tag{5.55}$$

As explained in the previous section, we can rewrite the time, t, in units of numbers of substitutions, d, as $t = d/3\lambda$. By plugging this expression into Equation (5.55), we obtain

$$p(d) = \frac{3}{4} - \frac{3}{4}e^{-\frac{4}{3}d}. \tag{5.56}$$

We note here that $0 \leq p(t) \leq 3/4$, since $0 \leq d \leq \infty$.

Now, we rearrange the Equation (5.56) and obtain:

$$d = -\frac{3}{4}\log\left(1 - \frac{4}{3}p(d)\right). \tag{5.57}$$

We aim to obtain an estimator for d, $\hat{d}$. Before estimating d, we obtain an estimate for $p(t)$, $\hat{p}$, based on the method of moments. We use the two sequences as data and count the number of segregating sites in the two sequences, x. The probability $p(t)$ is the probability of any site being segregating, meaning each site undergoes a Bernoulli trial (Box 20 on page 95) with success probability $p(t)$, and thus the expectation for a site to be segregating is $p(t)$. When comparing two sequences with n sites, we observe n Bernoulli trials (one for each site), and x/n is the expectation for a site to be segregating in this sample. In the method of moments, the estimate for the expectation $p(t)$ is the sample expectation, $\hat{p} = x/n$. Thus, our estimator for d based on the method of moments estimator for $p(t)$ is:

$$\hat{d} = -\frac{3}{4}\log\left(1 - \frac{4}{3}\frac{x}{n}\right). \tag{5.58}$$

Example: In Figure 5.8, we display two sampled sequences (the same as taxon 1 and 2 in Figure 5.7) together with their (typically unknown) ancestor. For the alignment of the two sampled sequences, we count $x = 2$ differences for the total length of $n = 8$ nucleotides. The estimated probability of substitution is, therefore, $\hat{p} = x/n = 2/8 = 0.25$. According to the JC69 pairwise distance formula just derived, the distance estimate is $\hat{d} = -\frac{3}{4}\log\left(1 - \frac{4}{3}\hat{p}\right) = -\frac{3}{4}\log\left(\frac{2}{3}\right) = 0.3$.

When estimating a parameter from data, the estimated parameter does not necessarily match the true value; thus, a parameter estimate is normally reported with a measure of uncertainty such as the variance or a confidence interval. There are multiple methods for calculating the variance of $\hat{d}$, one of which is the so-called *delta technique*. We refer the interested reader to Yang (2014, Appendix B), which describes this technique in detail. Below, we will explain how to obtain confidence intervals for the maximum likelihood method.

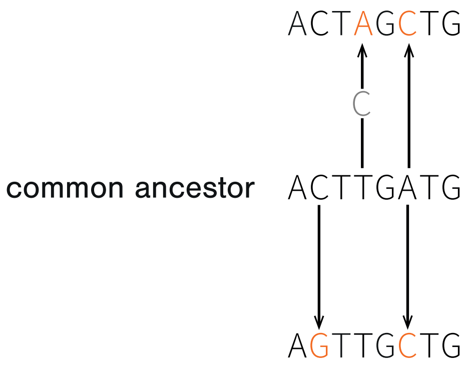


Figure 5.8: Example alignment of the two sequences ACTAGCTG and AGTTGCTG (taxon 1 and 2 in Figure 5.7) and their true (unknown) common ancestor sequence ACTTGATG. The arrows indicate substitutions, and the letter on one of the arrows indicates multiple substitutions.

5.4.3 Pairwise distances using a maximum likelihood approach for JC69

In this section, we derive the *maximum likelihood estimator (MLE)* for pairwise distances. Maximum likelihood estimators are explained in Box 25 on page 116. Further, in Box 26 on page 117, we explain the concept of confidence intervals (CI) and how to obtain confidence intervals for maximum likelihood estimators.

To derive the maximum likelihood estimator for the distance between a pair of sequences, we note that the probability of a substitution under the JC69 model in time t is $p = 3p_1(t)$, where $p_1(t) = \frac{1}{4} - \frac{1}{4}e^{-4\lambda t}$. The number of differences between two sequences of length n is binomially distributed:

$$P(x \text{ substitutions out of } n \text{ nucleotides}) = \binom{n}{x} p^x (1-p)^{n-x}. \tag{5.59}$$

After substituting p by $3p_1(t) = \frac{3}{4} - \frac{3}{4}e^{-4\lambda t} = \frac{3}{4} - \frac{3}{4}e^{-\frac{4}{3}d}$, Equation (5.59) is equal to

$$\binom{n}{x} \left(\frac{3}{4} - \frac{3}{4}e^{-\frac{4}{3}d}\right)^x \left(\frac{1}{4} + \frac{3}{4}e^{-\frac{4}{3}d}\right)^{n-x} = L(d; x), \tag{5.60}$$

which defines the likelihood function. Log-transformation of this leads to:

$$l(d; x) = \log \binom{n}{x} + x \log \left(\frac{3}{4} - \frac{3}{4}e^{-\frac{4}{3}d}\right) + (n-x) \log \left(\frac{1}{4} + \frac{3}{4}e^{-\frac{4}{3}d}\right). \tag{5.61}$$

Box 25: Maximum likelihood estimator (MLE)

Let $P(X = data|\theta)$ be a probability density of a random variable X given parameters θ. Then the function in θ

$$L(\theta; data) = P(X = data|\theta) \tag{B25.1}$$

is called the *likelihood function* or *likelihood* for short. For observed data $data$, the *maximum likelihood estimator (MLE)*, is the parameter value for which the likelihood is the highest,

$$\hat{\theta} = \text{argmax}_\theta L(\theta; data). \tag{B25.2}$$

Very often, the terms probability and likelihood are used interchangeably. However, they describe very different concepts in a strict mathematical sense; the probability is a function of data, while the likelihood is a function of model parameters.

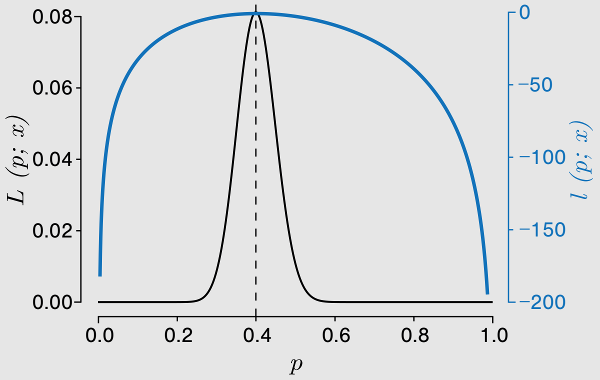
We illustrate the concept of MLE with the example of repeatedly rolling a 6-sided die. After we roll the die $n = 100$ times and observe $x = 40$ sixes, we want to estimate the parameter "probability of rolling a six", denoted by p, using an MLE. Let the random variable X denote the number of sixes out of n die rolls. Then $x = 40$ is one realisation of this random variable, our data are $data = x$, and our parameter is $\theta = p$. The random variable X is binomially distributed, $P(X = x|p) = \binom{n}{x} p^x (1-p)^{n-x}$. Thus, the likelihood function in our experiment is

$$L(p; x) = P(X = x|p) = \binom{100}{40} p^{40} (1-p)^{60}. \tag{B25.3}$$

The MLE is the value of p that best explains the observed data, that is, the value that maximises Equation (B25.3). A necessary condition for minima and maxima is that the first derivative equals 0. As we are only interested in the value at which the likelihood function takes its maximal value, the likelihood function can be transformed with functions that do not move the location but only the actual value of the maximum, such as taking the logarithm. This specific transformation is called the log-likelihood function, $l(p; x)$:

$$l(p; x) = \log L(p; x) \overset{B25.3}{=} \log \binom{100}{40} + 40 \log p + 60 \log(1-p). \tag{B25.4}$$

As the log-likelihood function from our example shows, the logarithm changes multiplications to sums, thus simplifying getting an analytical solution to the maximisation. Below, we show the likelihood (black) and log-likelihood (blue) functions for the die experiment.



To find the location of the maximum of the log-likelihood function, we calculate the first derivative of the log-likelihood function with respect to p and set the derivative to 0. For our example it is $\hat{p} = x/n = 40/100 = 0.4$. Since the second derivative is < 0, we conclude this is a maximum (as seen in the plot above).

Box 26: Confidence interval

A *confidence interval (CI)* is a measure of the uncertainty around a particular estimate. The interval is an estimate itself and depends on the observed realisation of a random experiment. The interval estimate is called a $(1-\alpha) \times 100\%$ confidence interval (e.g. a 95% confidence interval) if the true parameter ϑ lies within the estimated interval in $(1-\alpha) \times 100\%$ of the repeated random experiments. Note that if ϑ is a vector (e.g. $\vartheta = (\mu, \sigma^2)$ in a normal distribution), its confidence measure is called a *confidence region*.

The likelihood framework offers an easy estimate of the CI. For given data, the MLE (see Box 25 on page 116) is denoted with $\hat{\theta}$, and the *log-likelihood ratio function (LR)* is defined as:

$$\mathrm{LR}(\hat{\theta}, \vartheta) = 2(l(\hat{\theta}; data) - l(\vartheta; data)) = \log \left(\frac{L(\hat{\theta}; data)}{L(\vartheta; data)} \right)^2. \tag{B26.1}$$

The right side of Equation (B26.1) demonstrates why this function is called the log-likelihood ratio function.

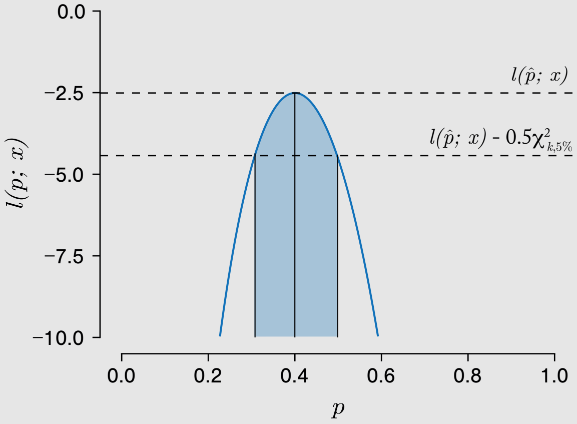
When data are generated under ϑ, we can apply Wilk's theorem (see also Section 7.2.1) to show that $\mathrm{LR}(\hat{\theta}, \vartheta)$ approximately follows a χ^2_k distribution (see Box 11 on page 78):

$$\mathrm{LR}(\hat{\theta}, \vartheta) \sim \chi^2_k. \tag{B26.2}$$

In this case, the degree of freedom k corresponds to the length of the parameter vector ϑ.

The $(1-\alpha) \times 100\%$ CI is the set of parameter values θ where $\mathrm{LR}(\hat{\theta}, \theta) \le \chi^2_{k,\alpha}$, meaning all θ which are not in the tail of the χ^2_k distribution are in the CI. Each value θ in the $(1-\alpha) \times 100\%$ CI is a candidate for being the true ϑ, and in $\alpha \times 100\%$ of cases, ϑ is not contained inside the $(1-\alpha) \times 100\%$ confidence interval.

Example: To calculate the 95% confidence interval of the parameter $\theta = p$ for the die throwing example from Box 25 on page 116, we first calculate the value of $l(\hat{p}; x)$. Then, we look up the value of $\chi^2_{k,5\%}$ in the χ^2-table, and calculate the values for ϑ such that $l(\vartheta; x) > l(\hat{p}; x) - 0.5\chi^2_{k,5\%}$:



Based on this procedure, we can determine the 95% confidence interval for the probability of rolling a six in our example. We rolled the die 100 times and observed 40 sixes, corresponding to the probability of $\hat{p} = 0.4$. The 95% confidence interval is $[0.308, 0.499]$ (shown in blue).

Note that if the die were fair, we would expect $p = 1/6$, which is not within the 95% CI of this particular realisation.

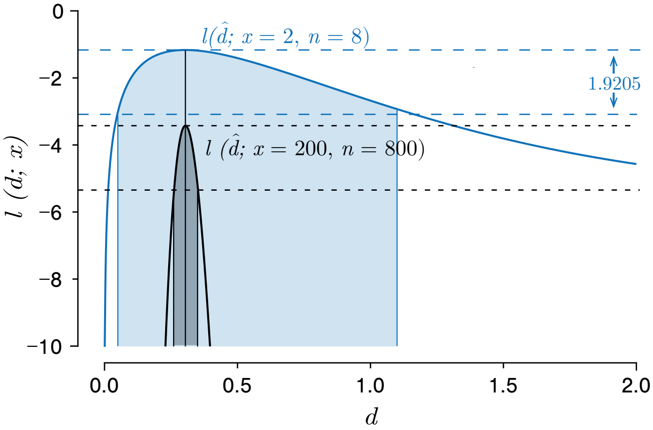


Figure 5.9: Maximum likelihood estimation of the distance between two nucleotide sequences of length 8 (in blue) and of length 800 (in black). The log-likelihood $l(d;x)$ is shown on the y-axis. The sequence distance d is shown on the x-axis. The maximum likelihood distance estimate is the x-value at which the curve reaches a maximum. The 95% confidence intervals are indicated with blue and black shaded areas.

To obtain the maximum likelihood estimate of the distance under JC69, we need to differentiate the Equation (5.61) with respect to d and set the derivative to 0. We then obtain $\hat{d} = -\frac{3}{4}\log\left(1 - \frac{4x}{3n}\right)$, which is the same as the method of moments estimator under JC69 we derived in Section 5.4.2.

Let us calculate the MLE and the CI of the distance under JC69 for the two sequences shown in Figure 5.8 (assuming we do not know the ancestral sequence). There are two differences between the 8-nucleotide long sequences, so the maximum likelihood distance estimate is $\hat{d} = -\frac{3}{4}\log\left(1 - \frac{4\times 2}{3\times 8}\right) = 0.3$. The CI for this estimate is $[0.05, 1.17]$ (blue shaded area in the Figure 5.9). The CI is not symmetric around the MLE, and the uncertainty in the distance estimate is quite large (wide CI). This is due to the small amount of information our sequences carry because of their short length. If we had more data in the form of longer sequences, we would have more confidence in our estimate. In fact, when we repeat the calculations for sequences of length 800 instead of 8, and with 200 differences instead of 2, we obtain the same MLE of $\hat{d} = -\frac{3}{4}\log\left(1 - \frac{4\times 200}{3\times 800}\right) = 0.3$, but $\text{CI} = [0.26, 0.35]$. As expected, the CI is much more narrow than in the short sequence example — see the black shaded area in Figure 5.9.

In Table 5.2, we provide the distances for other nucleotide substitution models that can be derived as described here.

5.5 Allowing for rate variation across sites

So far, we have only considered models where all sites in the sequence evolve under the same model. However, this may not always be a reasonable assumption because the substitution rates might vary between different sites. This variability can be due to variable mutation rates in different parts of the genome (e.g. the polymerase could have different error rates across different parts of the genome) or due to variable selective pressure on different parts of the phenotype (e.g. a viral sequence could have parts that are under strong selective pressure to escape the host immune system and other parts under strong pressure to remain the same to allow the virus to use conserved host receptors to enter the host cells). Thus, we extend the substitution rate models to account for this variability, referred to as *rate heterogeneity*.

The extension of variable rates across different sites in the sequence is often modelled by replacing constant rates with a Γ distributed random variable (see Box 14 on page 81) for each site (denoted JC69+Γ, K80+Γ, and so on). Typically, a $\Gamma(\alpha, \alpha)$ distribution is chosen. This distribution has mean 1, meaning that the average substitution rate remains the same as in the original model. When considering the rates for all sites in a sequence of length n, the empirical rate distribution corresponds to n draws from the Γ distribution.

For the JC69 evolutionary model with Γ distributed rates, we replace the single λ parameter with λR where R is a $\Gamma(\alpha, \alpha)$ distributed random variable. According to Equation (B14.3), the chosen Γ distribution has mean 1, and thus $\mathrm{E}(\lambda R) = \lambda$. The transition probability then becomes $p(t) = \frac{3}{4} - \frac{3}{4}e^{-4\lambda R t} = \frac{3}{4} - \frac{3}{4}e^{-\frac{4}{3}dR}$.

For a particular site, we do not know which value the random variable R takes. Thus, we average over all possible values for R to obtain the expected transition probability:

$$\mathrm{E}(p) = \int_0^\infty (\frac{3}{4} - \frac{3}{4}e^{-\frac{4}{3}dr}) f_R(r; \alpha, \alpha)\, \mathrm{d}r = \frac{3}{4} - \frac{3}{4}\left(1 + \frac{4d}{3\alpha}\right)^{-\alpha}. \tag{5.62}$$

To calculate the pairwise sequence distance under this extended model, we again (as in Section 5.4.2) equate the observed proportion of different sites between the two sequences x/n (the sample expectation for a site to be segregating) to the expectation of a site to be segregating, $\mathrm{E}(p)$, to get $\hat{d} = \frac{3}{4}\alpha\left(\left(1 - \frac{4}{3}\hat{p}\right)^{-1/\alpha} - 1\right)$. Furthermore, we can use the maximum likelihood framework, as was done in Section 5.4.3, to derive maximum likelihood distances.

Let us look back at our example alignment from Figure 5.8 (again ignoring the ancestral sequence). The length of the two sequences is $n = 8$ nucleotides, of which $x = 2$ sites are different. We obtained a maximum likelihood estimate of 0.3 for the distance under JC69 without site variation. We take the site variation into account by assuming $\Gamma(2, 2)$ distributed substitution rates. This results in an estimated distance of

$$\hat{d}_{\mathrm{JC69+\Gamma}} = \frac{3}{4}\alpha\left(\left(1 - \frac{4}{3}\hat{p}\right)^{-1/\alpha} - 1\right) = 0.34 > 0.3 = \hat{d}_{\mathrm{JC69}} \tag{5.63}$$

between the two sequences. The distance estimate with site variation is bigger than the one we obtained when considering a simple JC69 model. Therefore, in our example, ignoring the site variation — given it is present — leads to underestimating sequence distance. This observation holds in general, as we will show in Theorem 5.5.1.

Theorem 5.5.1. *In the JC69 model, not modelling among-site rate variation leads to a smaller sequence distance estimate compared to assuming a* Γ *distributed among-site rate variation.*

Proof. We derived the distance estimator for JC69, $\hat{d}_{\mathrm{JC69}} = -\frac{3}{4}\log\left(1-\frac{4}{3}\hat{p}\right)$ and the estimator for JC69+Γ, $\hat{d}_{\mathrm{JC69+\Gamma}} = \frac{3}{4}\alpha\left(\left(1-\frac{4}{3}\hat{p}\right)^{-1/\alpha}-1\right)$. To prove this theorem, we need to prove that

$$-\frac{3}{4}\log\left(1-\frac{4}{3}\hat{p}\right) \leq \frac{3}{4}\alpha\left(\left(1-\frac{4}{3}\hat{p}\right)^{-1/\alpha}-1\right) \tag{5.64}$$

for all $\alpha > 0$ and $0 \leq \hat{p} < 3/4$.

We transform Equation (5.64) by multiplying both sides by $4/3\alpha$, applying $a\log x = \log x^{a}$, and exponentiating both sides. This means that proving Equation (5.64) is equivalent to proving

$$\left(1-\frac{4}{3}\hat{p}\right)^{-1/\alpha} \leq \exp\left(\left(1-\frac{4}{3}\hat{p}\right)^{-1/\alpha}-1\right). \tag{5.65}$$

We define $x = \left(1-\frac{4}{3}\hat{p}\right)$. As $0 \leq \hat{p} < 3/4$, x ranges between 0 and 1, $0 < x \leq 1$. We now expand the right side of Equation (5.65) using Equation (B16.1) of the exponential function (Box 16 on page 88):

$$\begin{aligned}\exp\left(\left(1-\frac{4}{3}\hat{p}\right)^{-1/\alpha}-1\right) &= \exp\left(x^{-1/\alpha}-1\right) = \sum_{n=0}^{\infty}\frac{\left(x^{-1/\alpha}-1\right)^{n}}{n!}\\ &= 1+\left(x^{-1/\alpha}-1\right)+\sum_{n=2}^{\infty}\frac{\left(x^{-1/\alpha}-1\right)^{n}}{n!}\\ &= x^{-1/\alpha}+\sum_{n=2}^{\infty}\frac{\left(x^{-1/\alpha}-1\right)^{n}}{n!}.\end{aligned} \tag{5.66}$$

Because $0 < x \leq 1$, it follows that $x^{-1/\alpha} = \frac{1}{x^{1/\alpha}} \geq 1$ and thus $(x^{-1/\alpha}-1) \geq 0$. We can then conclude

$$\exp\left(x^{-1/\alpha}-1\right) = x^{-1/\alpha} + \underbrace{\sum_{n=2}^{\infty}\frac{\left(x^{-1/\alpha}-1\right)^{n}}{n!}}_{\geq 0} \geq x^{-1/\alpha}. \tag{5.67}$$

Per definition $x = \left(1-\frac{4}{3}\hat{p}\right)$, we obtain Equation (5.65), which proves that the JC69 distance is always less than or equal to the JC69+Γ distance. □

For more complex substitution models such as the GTR model, direct integration of Equation (5.62) is impossible. Thus, when used for practical inference, this integration is handled using an approximation (Yang 1996), in which the rate at a given site is assumed to be distributed across a small and fixed number of discrete values (rate categories) corresponding to equal-probability quantiles of the $\Gamma(\alpha, \alpha)$ distribution. The number of discrete rate categories is usually displayed as a subscript of the Γ distribution; for example, the K80 model with 4 discrete rate categories will be written as K80+Γ_4. Numerically averaging over these discrete rate values approximates integrating over the Γ distribution.

Table 5.2 lists a collection of distance estimators for several substitution models, both with and without site rate heterogeneity. Not all substitution models are present in this list (GTR and UNREST are missing, for example), as these models lack a closed-form solution for the transition probability function, which is used to derive the distance estimators in this list.

As discussed by Felsenstein (2003), another common approach to dealing with site-to-site heterogeneity is simply allowing sites to be either "variable" or "invariable", with mutation strictly forbidden at "invariable" sites. This can be done by introducing a parameter p_{inv} to represent the probability of any given site being invariable, which then serves a roughly analogous role to the shape parameter in the $\Gamma(\alpha, \alpha)$ distribution. The idea behind this approach is to allow portions of the genome — perhaps those under strong selection — to remain fixed without influencing the inferred substitution rate for neutral sites. Substitution models with this modification are sometimes given the suffix +I (e.g. GTR+I).

In addition to rate variation across sites, we may have rate variation across branches. We touch upon that topic in Section 6.4.3.

5.6 Amino acid substitution models

In the previous section, we focused on the evolution of nucleotides and the models that quantify nucleotide substitution rates. However, selection takes place at the level of the phenotype. This section focuses on studying evolution at this level and quantifying it using amino acid substitution models.

Example: The human immunodeficiency virus (HIV) is a persistent infection that quickly adapts to ever-changing immunological environments. The envelope protein is the only protein that sticks out of the viral membrane and is visible to the immune system. As soon as an individual becomes infected with HIV, the immune system starts producing antibodies directed at the envelope protein on the surface of the HIV virions. Thus, the immune system imposes selective pressure on the virus population, leading to immune escape. The adaptation of the amino acid sequence of the HIV envelope protein in response to the antibody response of the host immune system is an example of evolution acting at the level of phenotype — see Figure 5.10.

Model	Distance estimator	
JC69 (Jukes and Cantor 1969)	$\hat{d}_{\mathrm{JC69}} = -\frac{3}{4}\log\left(1-\frac{4}{3}\hat{p}\right)$, where $\hat{p} = {}^{x}/_{n}$, x is the number of segregating sites, n is the sequence length.	(5.68) (5.69)
K80 (Kimura 1980)	$\hat{d}_{\mathrm{K80}} = -\frac{1}{2}\log(1-2S-V) - \frac{1}{4}\log(1-2V)$, where S is the proportion of sites with transitional differences, V is the proportion of sites with transversional differences.	(5.70)
HKY (Hasegawa, Yano and Kishino 1984)	$\hat{d}_{\mathrm{HKY}} = 2\left(\frac{\pi_\mathsf{T}\pi_\mathsf{C}}{\pi_\mathsf{T}+\pi_\mathsf{C}} + \frac{\pi_\mathsf{A}\pi_\mathsf{G}}{\pi_\mathsf{A}+\pi_\mathsf{G}}\right)a - 2\left(\frac{\pi_\mathsf{T}\pi_\mathsf{C}(\pi_\mathsf{A}+\pi_\mathsf{G})}{\pi_\mathsf{T}+\pi_\mathsf{C}} + \frac{\pi_\mathsf{A}\pi_\mathsf{G}(\pi_\mathsf{T}+\pi_\mathsf{C})}{\pi_\mathsf{A}+\pi_\mathsf{G}} - (\pi_\mathsf{T}+\pi_\mathsf{C})(\pi_\mathsf{A}+\pi_\mathsf{G})\right)b$, where $a = -\log\left(1 - \frac{S}{2\left(\frac{\pi_\mathsf{T}\pi_\mathsf{C}}{\pi_\mathsf{T}+\pi_\mathsf{C}} + \frac{\pi_\mathsf{A}\pi_\mathsf{G}}{\pi_\mathsf{A}+\pi_\mathsf{G}}\right)} - \frac{\left(\frac{\pi_\mathsf{T}\pi_\mathsf{C}(\pi_\mathsf{A}+\pi_\mathsf{G})}{\pi_\mathsf{T}+\pi_\mathsf{C}} + \frac{\pi_\mathsf{A}\pi_\mathsf{G}(\pi_\mathsf{T}+\pi_\mathsf{C})}{\pi_\mathsf{A}+\pi_\mathsf{G}}\right)V}{2\left(\pi_\mathsf{T}\pi_\mathsf{C}(\pi_\mathsf{A}+\pi_\mathsf{G}) + \pi_\mathsf{A}\pi_\mathsf{G}(\pi_\mathsf{T}+\pi_\mathsf{C})\right)}\right)$, $b = -\log\left(1 - \frac{V}{2(\pi_\mathsf{T}+\pi_\mathsf{C})(\pi_\mathsf{A}+\pi_\mathsf{G})}\right)$, S is the proportion of sites with transitional differences, V is the proportion of sites with transversional differences.	(5.71) (5.72) (5.73)
TN93 (Tamura and Nei 1993)	$\hat{d}_{\mathrm{TN93}} = \frac{2\pi_\mathsf{T}\pi_\mathsf{C}}{(\pi_\mathsf{T}+\pi_\mathsf{C})}(a_1 - (\pi_\mathsf{A}+\pi_\mathsf{G})b) + \frac{2\pi_\mathsf{A}\pi_\mathsf{G}}{(\pi_\mathsf{A}+\pi_\mathsf{G})}(a_2 - (\pi_\mathsf{T}+\pi_\mathsf{C})b) + 2(\pi_\mathsf{T}+\pi_\mathsf{C})(\pi_\mathsf{A}+\pi_\mathsf{G})b$, where $a_1 = -\log\left(1 - \frac{(\pi_\mathsf{T}+\pi_\mathsf{C})S_1}{2\pi_\mathsf{T}\pi_\mathsf{C}} - \frac{V}{2(\pi_\mathsf{T}+\pi_\mathsf{C})}\right)$, $a_2 = -\log\left(1 - \frac{(\pi_\mathsf{A}+\pi_\mathsf{G})S_2}{2\pi_\mathsf{A}\pi_\mathsf{G}} - \frac{V}{2(\pi_\mathsf{A}+\pi_\mathsf{G})}\right)$, $b = -\log\left(1 - \frac{V}{2(\pi_\mathsf{T}+\pi_\mathsf{C})(\pi_\mathsf{A}+\pi_\mathsf{G})}\right)$, S_1 is the proportion of sites with two different pyrimidines (T, C), S_2 is the proportion of sites with two different purines (A, G), V is the proportion of sites with transversional differences.	(5.74) (5.75) (5.76) (5.77)
JC69+Γ	$\hat{d}_{\mathrm{JC69+\Gamma}} = \frac{3}{4}\alpha\left(\left(1-\frac{4}{3}\hat{p}\right)^{-1/\alpha} - 1\right)$, where α is the shape and the rate of the Γ distribution.	(5.78)
K80+Γ	$\hat{d}_{\mathrm{K80+\Gamma}} = \frac{1}{2}\alpha\left((1-2S-V)^{-1/\alpha} - 1\right) + \frac{1}{4}\alpha\left((1-2V)^{-1/\alpha} - 1\right)$, where α is the shape and the rate of the Γ distribution, S is the proportion of sites with transitional differences, V is the proportion of sites with transversional differences.	(5.79)

Table 5.2: Distance estimators from Yang (2014) for different nucleotide substitution models.

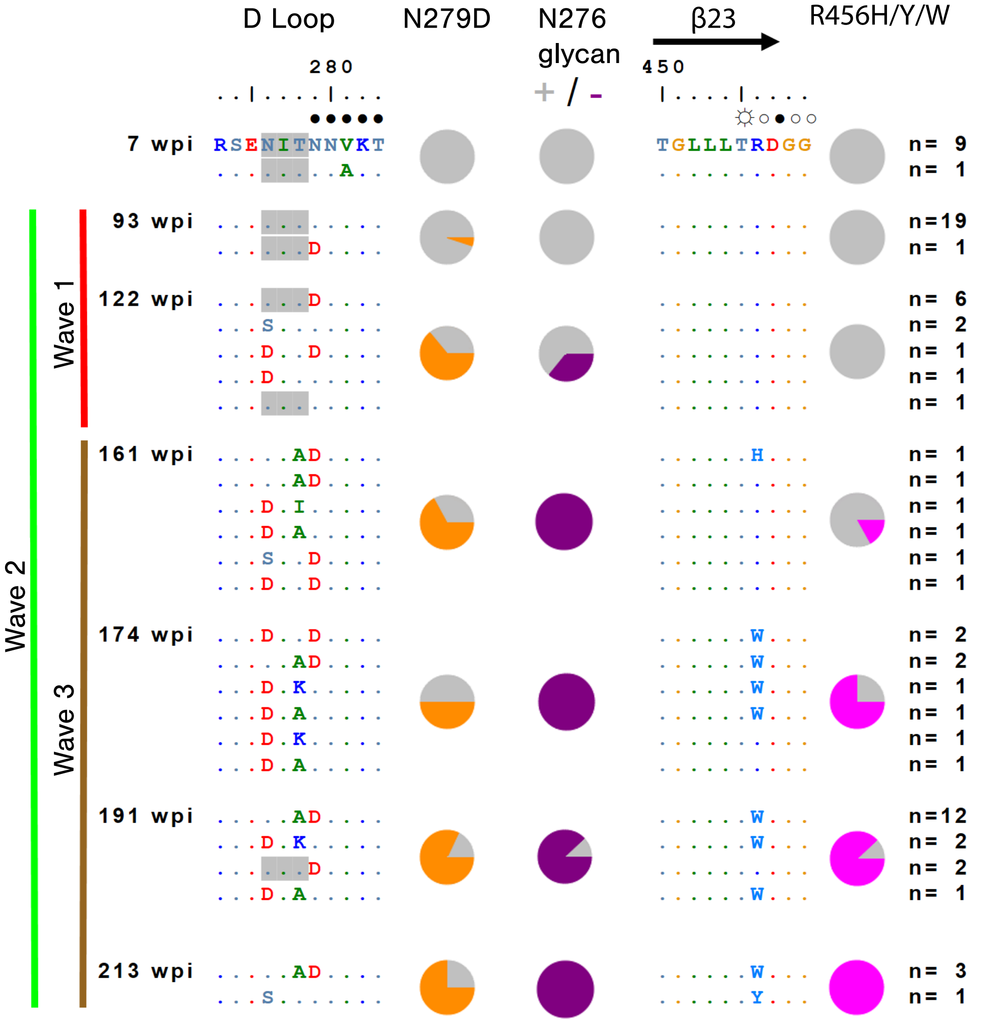


Figure 5.10: Adaptation of the HIV envelope protein (the protein on the surface of HIV virions) to the selective pressure exerted by the immune system of the host over time. Viral strains were isolated over 7 weeks post-infection (wpi) up to 213 wpi. Small parts of the viral sequence are shown in this figure, namely the D loop and β23. In the patient, three waves of antibodies were identified, each directed against a different target. Their timing is summarised on the left. Three specific mutations (N279D, N276 glycan, R456H/Y/W) were responsible for the failure of the antibodies. The pie charts show the presence of these mutations in the isolated viral strains at different time points. Figure adapted from Wibmer et al. (2013).

The amino acid composition of the envelope protein sequences in Figure 5.10 changes over time (the weeks post-infection (wpi) are displayed on the left of the figure). Three sites — 276, 279, and 456 — have a particularly significant impact on the success or failure of the autologous antibody response, which is the antibody response that follows a new mutation. The pie charts show how often mutations were found at these sites in the sampled viral sequences. Generally, we observe that the fraction of viral strains bearing mutations at these sites increases. This can be seen as a hint of selection for viral variants increasingly resistant to the autologous response.

This example nicely illustrates viral evolution at the phenotypic level. As evolution is the result of mutation and selection, we are interested in whether we can quantify the substitution rate of the amino acids and detect the presence of selection. To estimate substitution rates, we will in the following extend the substitution models presented earlier. To discover hints of selection, we then derive statistical tests (Section 5.7.2).

5.6.1 Definition of amino acid substitution models

We now define a substitution model at the amino acid level. Again, the process underlying the amino acid substitutions is modelled as a Markov chain with all the properties mentioned before. However, instead of four nucleotides, there are 20 amino acids. Thus, the state space of the Markov model is 20-dimensional, and the substitution rate matrix Q — and with it, the transition probability matrix $P(t) = e^{Qt}$ — is a 20×20 matrix with, in the most general case, 380 parameters (the diagonal is again chosen such that the row sum is 0).

The *amino acid substitution models* are more difficult to configure than the nucleotide substitution models. Like in nucleotide models, the Q matrix is ideally defined such that it ensures the time-reversibility of the model, as otherwise, downstream statistical analyses will be difficult (see Theorem 5.3.3 for the structure of Q if it defines a time-reversible process).

Typical datasets do not contain enough information to allow us to estimate hundreds of parameters (time-reversible models may have more than 200 parameters). Thus, the entries of amino acid substitution matrices are specified empirically or mechanistically, in contrast to nucleotide models, where entries are usually estimated. One can use empirical substitution rates (fixed rates that were previously suggested for the particular biological system) to fill the matrix. Alternatively, a probabilistic model can be used, which takes into account the properties of the individual amino acids and how easy it is to change from one amino acid to another, either from a chemical or a codon point of view.

Some of the most often used empirical amino acid substitution models can be roughly divided into two categories. The first category is the amino acid substitution scoring matrices, originally derived for scoring sequence alignments. Some of the most well-known models are the PAM matrices (point accepted mutation) introduced by Dayhoff, Schwartz and Orcutt (1978), the JTT model by Jones, Taylor and Thornton (1992), and the BLOSUM substitution matrices (BLOSUM62 mentioned briefly in Section 3.1.5 and shown on Figure 3.7) introduced by Henikoff and Henikoff (1992). The PAM matrices were derived from related protein sequences (above 85% similarity) using a parsimony-based criterion. JTT used a similar counting approach based on a much larger protein database than PAM. The BLOSUM matrices were calculated from more divergent sequences using existing alignments with different similarity percentages. For example, BLOSUM62 (Figure 3.7) was calculated based on alignments that share at least 62% similarity among sequences.

All models in this category are not based on explicit mechanistic substitution models but rather on observed substitution patterns. In particular, PAM and BLOSUM matrices were originally designed for scoring alignments, but they encode the information about the transition probabilities. For example, the BLOSUM matrix entries are defined as

$$S_{i,j} = \frac{1}{\lambda}\log(\frac{p_{i,j}}{\pi_i \pi_j}), \tag{5.80}$$

where λ is a scaling factor set arbitrarily such that the values are integer, $p_{i,j}$ is the empirical substitution probability, and π_i and π_j are the empirical amino acid frequencies. Thus, in terms of models as described in Section 5.2, the matrices of these empirical models state the transition probability matrices $P(t)$ (with value transformations necessary for PAM and BLOSUM) rather than the substitution rate matrices Q.

The second category of models is based on a mechanistic view of amino acid substitution. Each model was first defined mechanistically as a continuous-time Markov chain, and its parameters were estimated using maximum likelihood from large protein databases. While these models are mechanistic, they still have around 200 parameters. This makes it impossible to estimate all the parameters from a single dataset, meaning that empirical rate values are often used, while the stationary frequencies are estimated for the dataset in question.

Examples of such models include WAG (Whelan and Goldman 2001) and LG (Le and Gascuel 2008). As different organisms may have very different underlying evolutionary mechanisms, it is also useful to have mechanistic models with parameters estimated from specific datasets. For example, Nickle et al. (2007) estimated a model specific to HIV-1 subtype B, including estimates for within-host and between-host substitution matrices, and Dang et al. (2010) estimated a model specific to influenza proteins. All of these models provide a substitution rate matrix Q that can be used in sequence distance estimation and phylogenetic tree reconstruction (as described in Chapter 6).

5.6.2 JC69-like distance estimation for amino acid sequences

An amino acid substitution model allows us to define a measure of distance between two sequences as we did for the nucleotide substitution models. We use the same procedure to estimate the distance between amino acid sequences with an empirical or mechanistic Q-matrix. In the simplest, JC69-like model of amino acid substitution, all substitutions have the same rate λ. Thus, the mean rate of substitution is 19λ. The expected time to substitution is

$\frac{1}{19\lambda}$. This translates to time $t = \frac{d}{19\lambda}$ between two amino acid sequences. The distance estimator between the two sequences is $\hat{d} = \frac{19}{20} \log\left(1 - \frac{20x}{19n}\right)$, where n is the length of the sequences and x the number of substitutions.

5.7 Codon substitution models

We will now discuss the definitions and properties of codon substitution models. These models allow us to estimate whether selection is acting on (parts of) the sequences.

5.7.1 Definition of codon substitution models

A *codon* consists of three nucleotides and encodes for one amino acid (see Figure 1.7). As there are four nucleotides, there are $4^3 = 64$ possible codons. However, during the translation of RNA into proteins, three *stop codons* — `TAA`, `TAG` and `TGA`, — stop the translation process. The codon substitution models disregard these three codons because any premature stop codons in the protein-coding sequence usually cause the sequence to be translated into a non-functional protein. Thus, the codon models account for transitions between 61 codons, resulting in a very large substitution rate and transition probability matrices (61×61 entries).

One of the codons, `ATG`, the so-called *start codon*, serves as the biological barcode, signalling that the protein code starts at that position. The start codon and the remaining 60 codons each encode an amino acid. However, there are only 21 amino acids, of which only 20 are physiologically relevant and appear in the genetic code. This means that several codons can encode the same amino acid. The "codon sun" in Figure 1.7 illustrates that some nucleotide substitutions do not lead to any changes at the phenotypic level and, thus, are less likely to be under selection.

Generally, each codon can change into nine other codons with one nucleotide substitution (see Figure 5.11 for an example using codon `CGG`). Nucleotide substitutions in the codon leading to the same amino acid are called *synonymous substitutions*, and nucleotide substitutions in the codon leading to a different amino acid are called *non-synonymous substitutions*. Codons that result from a nucleotide transition on the third position often translate to the same amino acid, and the codons that result from a transversion often produce a different amino acid (recall that transversions usually occur less frequently than transitions).

We will now introduce a codon model using notation following (Yang 2014). As nucleotide and amino acid models, the codon models are Markov chain models, but with 61 possible states. We denote codons in the codon models with capital letters (e.g. I, J) and nucleotides with small letters (e.g. i, j). Codon models assume that the rate of change from I to J is zero if I and J differ in more than one nucleotide position, meaning that only single nucleotide changes are allowed. In general, this means that we may assume arbitrary rates for the transition of each codon into the nine codons, which are one substitution away, which makes

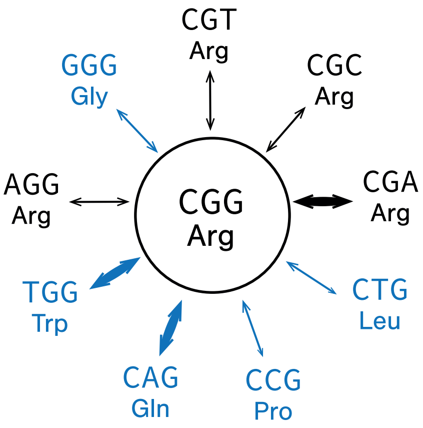


Figure 5.11: All possible codons into (or from) which the CGG codon could mutate with only one nucleotide substitution. The corresponding amino acid is shown as a three-letter code for each codon below. The non-synonymous substitutions are shown in blue. The synonymous substitutions are in black. The bigger arrows show transitions, and the smaller arrows show transversions. Figure inspired by Yang (2014).

$9 \times 61 = 549$ parameters. It is typically unfeasible to estimate that many parameters. Instead, some assumptions are made in the common codon models (Goldman and Yang 1994).

First, common codon models assume that the ratio of the synonymous transition rate to the synonymous transversion rate is the same across all codons I. Similarly, the ratio of the non-synonymous transition rate to the non-synonymous transversion rate is often assumed to be the same across all codons I. The associated parameters are κ for the transition/transversion ratio and ω for the non-synonymous/synonymous rate ratio. Furthermore, let π_I be the equilibrium frequency of codon I. We can assume that each codon frequency is a free parameter (all frequencies summing up to 1).

Each off-diagonal entry in the substitution rate matrix making these assumptions is then defined as (Nielsen and Yang 1998):

$$q_{IJ} = \begin{cases} 0 & \text{if I and J differ at more than 1 positions;} \\ \pi_J & \text{if I and J differ by a synonymous transversion;} \\ \kappa\pi_J & \text{if I and J differ by a synonymous transition;} \\ \omega\pi_J & \text{if I and J differ by a nonsynonymous transversion;} \\ \omega\kappa\pi_J & \text{if I and J differ by a nonsynonymous transition.} \end{cases} \tag{5.81}$$

In practice, these transition probabilities must be evaluated numerically. Due to the sizes of the transition matrices for the codon model, the complexity of the calculations is proportionately larger than those for the most general nucleotide substitution models (Felsenstein 2003).

5.7.2 Detecting selection: d_N/d_S ratio

The presence of selection acting on a nucleotide sequence can be revealed by comparing the amount of synonymous and non-synonymous nucleotide differences between two sequences. The idea behind this comparison is that if there are significantly more (or fewer) non-synonymous than synonymous differences than expected by chance, the protein was likely under selective pressure to specifically adapt its amino acid composition (or to remain unchanged). Otherwise, there was likely no selection acting on the protein.

Comparing non-synonymous and synonymous nucleotide differences between two sequences is a challenging task. We cannot compare the number of non-synonymous and synonymous differences directly because the probability of a random nucleotide substitution leading to a non-synonymous or synonymous change is not the same for each site (due to differing codon positions and the variable amounts of redundancy in codon encoding per amino-acid, see Figure 1.7). Thus, we have to scale these differences by the number of possible substitutions of the respective type.

The number of non-synonymous sites for two sequences is defined as the probability of a non-synonymous nucleotide change in either sequence times the sequence length[2]. Analogously, the number of synonymous sites for two sequences is the probability of a synonymous nucleotide change in either sequence times the sequence length. Thus, the sum of synonymous and non-synonymous sites for a pair of sequences equals the sequence length.

For two sequences, let us define d'_N as the ratio of the number of non-synonymous nucleotide differences to the number of non-synonymous sites. Similarly, let us define d'_S as the ratio of the number of synonymous nucleotide differences to the number of synonymous sites in our two sequences. Note that these definitions assume the absence of back-mutations (see Figure 5.7). Commonly used methods correct the d'_N and d'_S to account for this possibility, yielding d_N and d_S. In the upcoming section, we provide one way to estimate d_N and d_S.

The d_N/d_S *ratio* is typically reported, as it contains information on the abundance of selection between two nucleotide sequences. A ratio $d_N/d_S < 1$ means that non-synonymous substitutions happen less frequently than synonymous substitutions, often referred to as purifying selection (the genome is “purified” and substitutions are selected against; this corresponds to highly conserved phenotypes). A ratio $d_N/d_S > 1$ means that non-synonymous substitutions occur more frequently than synonymous substitutions, which leads to positive selection that accelerates the fixation of non-synonymous substitutions. A ratio $d_N/d_S = 1$ means no selection; synonymous and non-synonymous substitutions happen at the same rate.

[2] Note that this is an *effective* number of sites, allowing for the fact that changes at individual sites can be either synonymous or non-synonymous depending on the substituted character.

Several methods have been introduced for determining the d_N/d_S ratio (see Yang (2014) for an overview). Here, we will look at the counting method introduced by Nei and Gojobori (1986) to understand important concepts concerning the d_N/d_S ratio.

5.7.3 Counting method

We introduce the counting method for determining d_N/d_S (Nei and Gojobori 1986). We follow these three steps:

1. count the number of non-synonymous and synonymous differences between the two nucleotide sequences, referred to as N_d and S_d, respectively;
2. count the number of non-synonymous and synonymous sites in the two nucleotide sequences, referred to as N and S, respectively;
3. account for the unobserved nucleotide substitutions by applying the pairwise distance formula (such as Equation (5.58) when assuming the JC69 model) to the ratios $d'_N = {}^{N_d}/_{N}$ and $d'_S = {}^{S_d}/_{S}$ to obtain dN and dS.

We illustrate the counting method with an example, looking at the two sequences TTTCCTCCTCCT and TTCCAGCCTCCT (example from Yang (2014, Section 2.5); this example is used here with permission of the author), which each can be divided into four codons:

	codon 1	codon 2	codon 3	codon 4
sequence 1	TTT	CCT	CCT	CCT
sequence 2	TTC	CAG	CCT	CCT

From the codon sun (Figure 1.7) we can see that codons TTT and TTC encode F (phenylalanine), CCT encodes P (proline), and CAG encodes Q (glutamine). Thus, sequence 1 encodes the amino acid sequence FPPP and sequence 2 encodes the amino acid sequence FQPP, giving us a synonymous nucleotide change in codon 1 and a non-synonymous nucleotide change in codon 2.

Step 1: counting the number of non-synonymous and synonymous differences To calculate N_d and S_d, we consider each codon position I of the two aligned sequences separately and count the number of non-synonymous nucleotide differences N_d^I and the number of synonymous nucleotide differences S_d^I. When counting, we assume that only one nucleotide may change in each time step, resulting possibly in different substitution pathways. For the whole sequence, we have $N_d = \sum_{I=1}^{K} N_d^I$ and $S_d = \sum_{I=1}^{K} S_d^I$, where K is the number of codons in each sequence.

If the two codons at position I have the same nucleotide, we have $N_d^I = S_d^I = 0$. If the two codons only differ in one nucleotide, we have either $N_d^I = 1$ and $S_d^I = 0$, or $N_d^I = 0$ and $S_d^I = 1$. If the two codons differ in two positions, we know that $N_d^I + S_d^I = 2$. However, we have two possible orderings for accumulating nucleotide substitutions (two possible evolutionary pathways), which may lead to different N_d^I and S_d^I in each pathway. If the two codons differ in all three nucleotide positions, we have six possible pathways, and we only know that for all of them, $S_d^I + N_d^I = 3$ holds. If we have more than one possible pathway, we average the resulting N_d^I and S_d^I over the possible pathways (giving each pathway equal weight).

Example: To illustrate this procedure, we now determine N_d and S_d for our example. We start by looking at the first codon. In sequence 1, this is TTT, and in sequence 2, it is TTC. Both codons encode the same amino acid phenylalanine (F). Thus, we count one synonymous nucleotide substitution and no non-synonymous nucleotide substitutions. Codon 2 differs in two sites between sequences 1 and 2. We assume that only one nucleotide substitution can occur per time step, so we have two nucleotide substitution pathways that could account for the two changes. We need to average over the two possible ways:

pathway	S_d^2	N_d^2
CCT (P) → CAT (H) → CAG (Q)	0	2
CCT (P) → CCG (P) → CAG (Q)	1	1
average	0.5	1.5

This means that $S_d^2 = 0.5$ and $N_d^2 = 1.5$.

Codons 3 and 4 are the same, meaning no difference is counted.

We can now calculate the sum of all non-synonymous and synonymous nucleotide differences between the two sequences by summing up the differences in the different codon positions:

	codon 1		codon 2		codon 3		codon 4	
sequence 1	TTT		CCT		CCT		CCT	
sequence 2	TTC		CAG		CCT		CCT	
$N_d =$	0	+	1.5	+	0	+	0	= 1.5
$S_d =$	1	+	0.5	+	0	+	0	= 1.5

Step 2: counting the number of non-synonymous and synonymous sites Each codon consists of three nucleotides. Each nucleotide can change into one of the three other nucleotides. Thus, we list all possible single nucleotide substitutions for each codon of the original

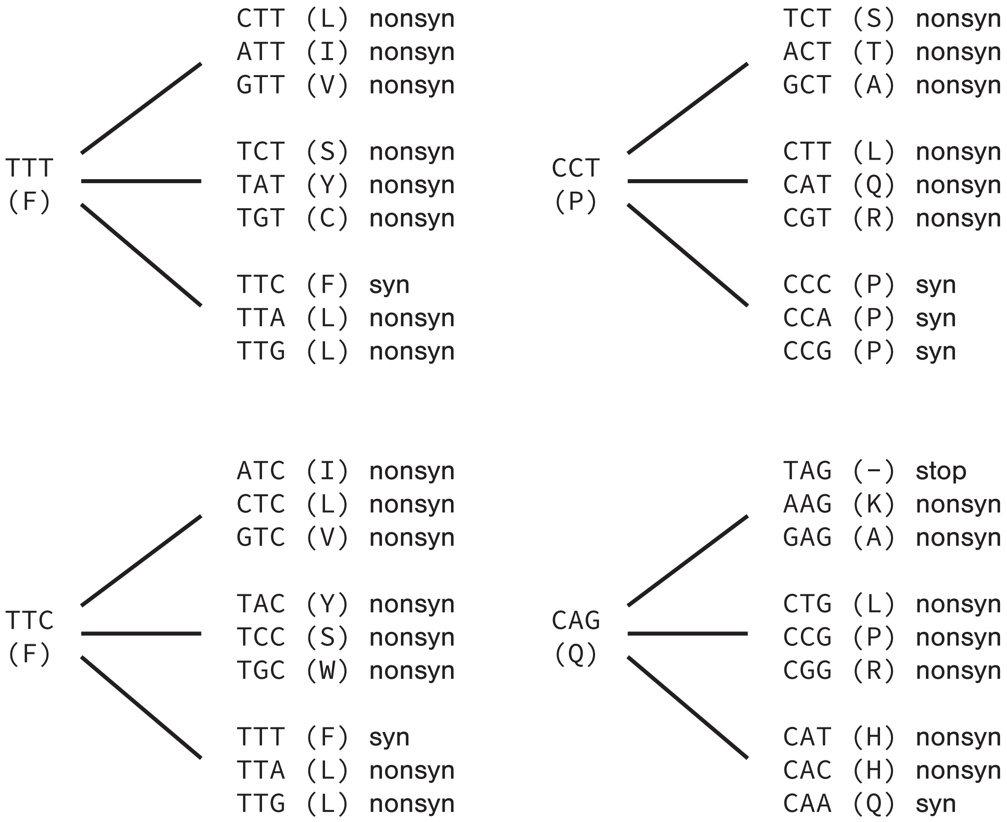


Figure 5.12: The four codons from our example and all codons reached from them by a single nucleotide change.

sequences and count whether this is a synonymous or non-synonymous nucleotide substitution. Note that stop codons are not considered. Figure 5.12 shows all the possible codons reachable through a single nucleotide substitution by the four codons in our two example sequences.

The number of synonymous nucleotide sites per codon I (S^I) is defined as the number of nucleotide sites in a codon (3) times the probability of obtaining a synonymous codon upon a single nucleotide substitution. Likewise, the number of non-synonymous nucleotide sites per codon I (N^I) is the number of sites in a codon (3) times the probability of obtaining a non-synonymous codon upon a single nucleotide substitution. Thus, $N^I + S^I = 3$. The total number of non-synonymous and synonymous sites for a sequence is obtained by summing N^I and S^I over all codons. The average of the per-sequence numbers is the number of non-synonymous and synonymous sites, N and S, for a pair of sequences.

Example: We determine the number of non-synonymous. and synonymous sites, N and S, in our example. The codon TTT has nine possible codons it can mutate into with one single nucleotide change. The substitution is synonymous only when the T at the third site of the codon mutates into a C. Thus, 1 out of 9 possible nucleotide substitutions are synonymous. According to the definition of the number of synonymous sites per codon, we need to multiply this number by 3; thus, codon TTT has $^1/_3$ synonymous sites. 8 out of 9 possible point substitutions lead to non-synonymous substitutions. Thus there are $3 \times {}^8/_9 = {}^8/_3$ non-synonymous sites in codon TTT. Following the same logic, the codon TTC has $3 \times {}^1/_9 = {}^1/_3$ synonymous sites and $3 \times {}^8/_9 = {}^8/_3$ non-synonymous sites. The codon CCT has $3 \times {}^3/_9 = 1$ synonymous sites and $3 \times {}^6/_9 = 2$ non-synonymous sites.

The codon CAG is more complicated because the substitution from C to T on the first site leads to the stop codon TAG. Stop codons are not considered in the codon substitution models and the counting method, as they would lead to an incomplete protein, which is typically not functional anymore. From the remaining 8 codons, only one codon results from a synonymous substitution. Thus, the codon CAG has $3 \times {}^1/_8 = {}^3/_8$ synonymous sites and $3 \times {}^7/_8 = {}^{21}/_8$ non-synonymous sites.

We can now calculate the number of non-synonymous and synonymous sites by averaging over the two sequences:

	codon 1		codon 2		codon 3		codon 4	
sequence 1	TTT		CCT		CCT		CCT	
sequence 2	TTC		CAG		CCT		CCT	
non-synonymous sites								
sequence 1	$^8/_3$	+	2	+	2	+	2	$= 8.67$
sequence 2	$^8/_3$	+	$^{21}/_8$	+	2	+	2	$= 9.29$
average								$N = 8.98$
synonymous sites								
sequence 1	$^1/_3$	+	1	+	1	+	1	$= 3.33$
sequence 2	$^1/_3$	+	$^3/_8$	+	1	+	1	$= 2.71$
average								$S = 3.02$

Step 3: accounting for evolutionary history We could now compare $d'_N = {}^{N_d}/_N$ and $d'_S = {}^{S_d}/_S$ in order to assess the amount of selection. However, the possible evolutionary steps between the two sequences are not taken into account when just looking at these quantities. This is why Nei and Gojobori (1986) corrected these quantities using the distance formula based on the JC69 molecular evolution model (Equation (5.58)) and defined these distances

as d_N and d_S respectively:

$$d_N = -\frac{3}{4}\log\left(1 - \frac{4}{3}\frac{N_d}{N}\right), \tag{5.82}$$

$$d_S = -\frac{3}{4}\log\left(1 - \frac{4}{3}\frac{S_d}{S}\right). \tag{5.83}$$

Example: Using the values calculated above, we can compute the d_N/d_S ratio:

$$d_N/d_S = \frac{-\frac{3}{4}\log\left(1 - \frac{4}{3}\frac{N_d}{N}\right)}{-\frac{3}{4}\log\left(1 - \frac{4}{3}\frac{S_d}{S}\right)} = \frac{\log\left(1 - \frac{4}{3}\frac{1.5}{8.98}\right)}{\log\left(1 - \frac{4}{3}\frac{1.5}{3.02}\right)} = 0.23. \tag{5.84}$$

In our example, we obtain a hint of purifying selection because the d_N/d_S ratio is less than 1.

Overall, the counting method includes many simplifications. The JC69 distance formula is based on the assumption that every nucleotide substitution occurs at the same rate. In particular, we do not take into account differences in transition and transversion rates and other codon biases. The counting method was therefore extended in the literature (e.g. Li (1993), Pamilo and Bianchi (1993), Comeron (1995), Ina (1995) and Tzeng, Pan and Li (2004)). In addition, another class of models was introduced: the maximum likelihood methods (Goldman and Yang 1994), in which the d_N/d_S ratio is obtained based on a maximum likelihood estimator. As discussed in Box 26 on page 117, such methods have the advantage of directly providing a confidence interval.

6 Phylogenetic trees

Nothing in evolution makes sense except in the light of phylogeny.
(Jay M. Savage (1997))

The following three chapters are on phylogenetics. This chapter will discuss how phylogenetic trees are reconstructed based on genetic sequences. The following two chapters will then introduce methods that allow us to take reconstruction uncertainty into account and to understand genotypic and phenotypic evolutionary processes occurring on such phylogenetic trees.

In what follows, we will first provide examples of phylogenetic trees. Second, we will introduce mathematical notation for and properties of these phylogenetic trees. Third, we will discuss the approaches for reconstructing phylogenetic trees: phenetic, cladistic, and probabilistic approaches. The last type, probabilistic methods, can be employed in a maximum likelihood or a Bayesian statistical setting. This chapter will discuss the maximum likelihood approach in detail, while the Bayesian approach will be explained in Chapter 10. We end this chapter by highlighting insights into HIV, which were obtained directly from reconstructed phylogenetic trees. In later chapters, we will discuss applying statistical methods to these reconstructed trees to develop quantitative insights into the evolutionary and population dynamical (e.g. epidemiological) processes governing the population from which the samples were taken.

6.1 Introduction to phylogenetic trees

The 1837 notebook of Charles Darwin shows a sketch of a phylogenetic tree (Figure 1.3). The phylogenetic tree displays evolutionary relationships between different individuals. The tree starts with a single ancestor individual (1) in the past and ends with branches without descendants, also called *leaves* or *tips* (A-D). Tips represent sampled individuals, individuals of whom we have genetic (or other types of) information. An *internal node*, the intercept of several branches, represents the most recent common ancestor of the individuals represented by the tips descending this node. Each individual has precisely one ancestor.

Phylogenetic trees were first reconstructed to display the evolutionary relationships of species. Typically, a genetic sequence of one individual per extant species is used to reconstruct the phylogeny; thus, each tip corresponds to one of these extant species. An internal node in a

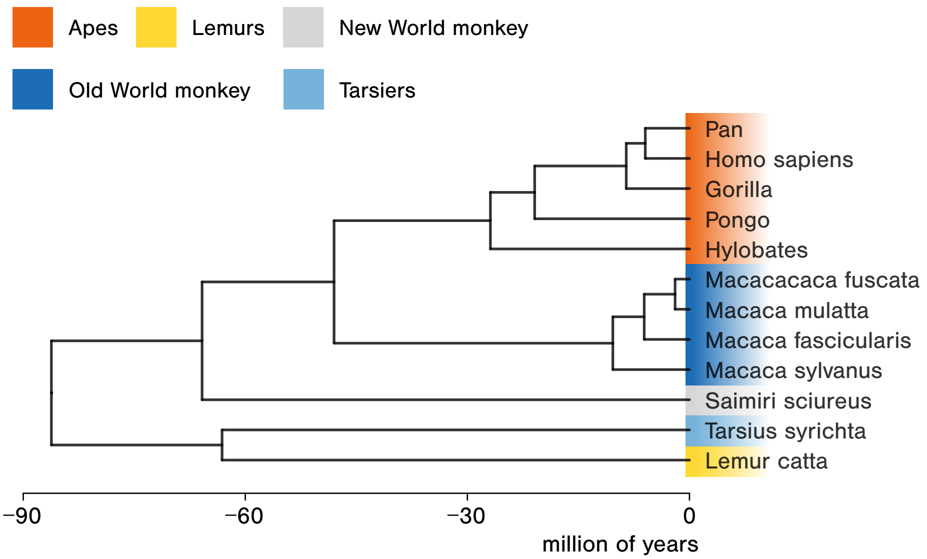


Figure 6.1: Phylogeny of primates. The tree was obtained based on results of the "Introduction to BEAST2" Taming the BEAST tutorial `https://taming-the-beast.org/tutorials/Introduction-to-BEAST2/`.

species tree represents a speciation event. In the primate phylogeny in Figure 6.1, we can see that humans are more closely related to chimpanzees (from genus *Pan*) than either humans or chimpanzees to gorillas. If lengths in calendar time units are assigned to the branches in a tree, one can read off speciation times. For example, we can see from the tree that humans and chimpanzees diverged around 6 million years ago. Or, based on the phylogenetic tree of mammals reconstructed by Bininda-Emonds et al. (2007) we can conclude that the most recent common ancestor of all mammals lived around 166 million years ago (note though that there is a lot of uncertainty and controversy around this estimate).

Hinchliff et al. (2015) reconstructed the tree of life containing 2.3 million species. Each tip in their tree actually represented around 1 000 extant species, meaning that subtrees were collapsed into tips. Additional genomic information helps to resolve the branching structure in the collapsed parts of the tree. The latest version of the tree can be explored with the web tool (Open Tree of Life (`https://tree.opentreeoflife.org/`)).

Sequences from different extant species can be sampled years apart. However, the evolution of species took millions of years; thus, sampling extant species within a few years' time can be interpreted as sampling at the same point in calendar time. This is visualised by putting all tips at the same point on a horizontal (time) axis in Figure 6.1.

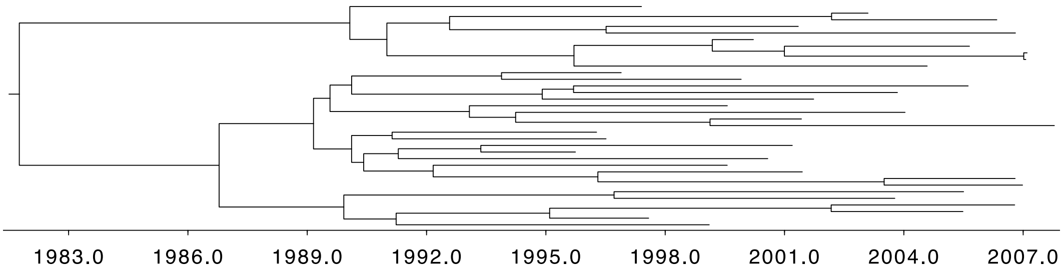


Figure 6.2: Phylogenetic tree of HIV samples from different patients created from the output file of the analyses in Stadler et al. (2011). The tree visualises an estimated Swiss HIV transmission cluster.

Over the past decade, the phylogenetic framework has also been heavily employed in the context of viral infectious diseases. To determine how a virus spreads in the host population, we need to obtain virus sequences from some infected individuals. The phylogeny of these viral sequences (one sequence per host, which typically is a consensus sequence, Section 3.3) is then used to approximate the *transmission tree*. As an example, consider the HIV phylogenetic tree in Figure 6.2. Viral samples of different patients form the tips of the tree. An internal node, in this context, represents the transmission of the pathogen from one host to another. Branch lengths in calendar time units can be interpreted as the time that has elapsed between transmission events. Compared to the phylogenetic tree of species, where branches represent millions of years of evolution, the phylogeny of pathogens covers a much shorter time period: usually months, years, or decades. Thus, in viral phylogenies, typically, not all tips are sampled at the same point in time; rather, the tips in the phylogeny of a virus can be sampled at different points in time throughout the epidemic. At the end of this chapter, we will discuss a number of insights into HIV that were obtained using such phylogenetic trees.

Throughout the COVID-19 pandemic, phylogenetic trees of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) reconstructed in real-time (Figure 6.3) became important tools to track the spread of the virus. Real-time means that the sequencing and reconstruction were done directly upon collection of the samples from patients. The unprecedented sequencing efforts around the world resulted in millions of sequences in the public domain (e.g. GenBank (`https://www.ncbi.nlm.nih.gov/genbank/`), ENA (`https://www.ebi.ac.uk/ena/browser/home`) and DDBJ (`https://www.ddbj.nig.ac.jp/index-e.html`)) and on the database GISAID (`https://gisaid.org/`), leading to the development of new phylogenetic tools to reconstruct trees on millions of sequences in real-time (e.g. see CoV2tree (`https://cov2tree.org/`); Turakhia et al. (2021)). Such advances allow policymakers to incorporate real-time phylogenetic information into their assessments.

Note that in all examples above, we assume that we include one sequence per species or one sequence per infected host (often a consensus sequence 3.3 of the pathogen). If we were to include several sequences per species, we would obtain trees connecting the individuals

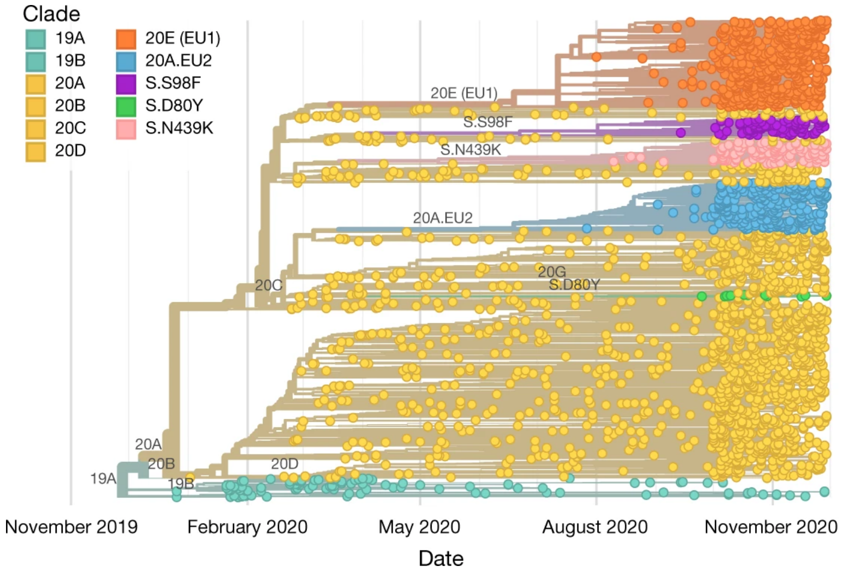


Figure 6.3: Real-time phylogenetic reconstructions from SARS-CoV-2 genomes were instrumental in informing public health policy during the COVID-19 pandemic. This figure, adapted from Hodcroft et al. (2021), shows a Nextstrain (Hadfield et al. 2018) maximum likelihood phylogeny of European SARS-CoV-2 genomes sequenced up until November 30, 2020.

within a species nested within the species tree (or, respectively, including several sequences per infected host will lead us to reconstruct trees connecting the different virions nested within the transmission tree). We will show such a tree in Section 6.6.4.

We will now introduce the mathematical notation of phylogenetic trees as well as important properties that will be required for the following sections on phylogenetic tree reconstruction and analysis.

6.2 The mathematics of phylogenetic trees

6.2.1 The mathematical definition of a phylogenetic tree

A *tree* consists of nodes and branches, with branches connecting the nodes such that no cycles are formed (otherwise, it would not be a tree, but rather a network, see Chapter 11). A node is of *degree* k if it has k branches attached. Here, we consider *binary trees*[1], which may be

[1] In the case of virion tracking or superspreading, we may require non-binary trees. See Section 9.4 for trees with multifurcations.

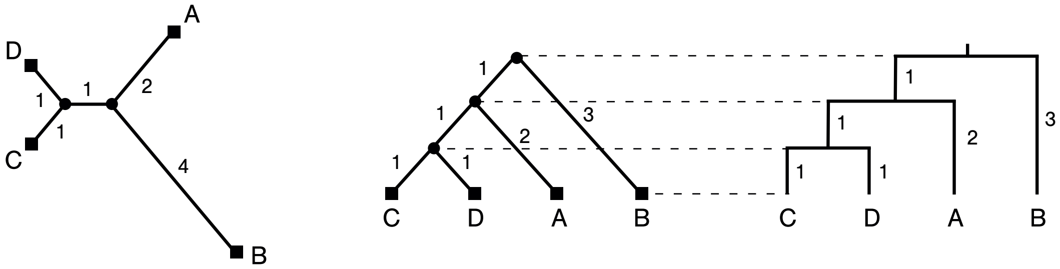


Figure 6.4: Unrooted phylogenetic tree on four tips (left) and a corresponding rooted tree (middle). The right tree is equivalent to the rooted tree in the middle but visualised as the trees in Figures 6.1 to 6.3.

unrooted or rooted. A binary *unrooted tree* (Figure 6.4, left) is defined as a tree with only degree-1 and degree-3 nodes. A degree-1 node is a *tip* of the tree, and a degree-3 node is an *internal node*. A binary *rooted tree* (Figure 6.4, middle) is an unrooted tree with one additional degree-2 node, this node is called the *root* of the tree[2]. A rooted tree can be obtained from an unrooted tree by dividing one of the branches into two by adding a new root node (in Figure 6.4 left, the branch leading to node B is divided by a root node to obtain the rooted tree shown in the middle). A *labelled tree* is a tree (rooted or unrooted) where each tip is assigned a unique label. In the trees shown in Figure 6.4, each branch has a specific length. A tree without branch lengths is called a *topology*. A *phylogenetic tree* is a rooted or unrooted labelled tree, with or without branch lengths. A rooted phylogenetic tree with branch lengths where all tips occur at the same time point in time is called an *ultrametric tree* (see e.g. Figure 6.1).

A branch attached to a tip is often called a *pendant branch* (in Figure 6.4, both phylogenies have four pendant branches corresponding to the four tips A, B, C, and D). Two tips whose adjacent pendant branches join in the same internal node are called a *cherry* (e.g. (C, D) and (A, B) in Figure 6.4, left and (C, D) in Figure 6.4, middle). A rooted tree containing a single cherry is called a *caterpillar tree* (Figure 6.4, middle). In a rooted tree, the set of tips descending from an internal node is also referred to as a *monophyletic group* or *clade*, for example, the tips (C, D, A) form a clade in Figure 6.4, middle.

Note that in the rooted trees in Figures 6.1 to 6.3, branching events are not represented by a single node that has an ancestral branch and two descendant branches attached. Instead, a branching event is displayed with the help of a line orthogonal to the branches from the root towards the tip. This is simply a different visualisation of rooted trees: the branches from the root to the tip are proportional to time, while the orthogonal lines display relationships. We provide an example of this visualisation in Figure 6.4, right.

[2] Alternatively, one can define a rooted tree as an unrooted tree where one of the degree-1 nodes is the origin of the tree (rather than a tip), and its direct descendant node is the root. Such trees are natural when considering birth-death models in Chapter 9

6.2.2 The Newick tree format

Most representations of phylogenetic trees are intended specifically for human inspection. However, when interacting with computer programs, it is often useful to have a more compact description format. One very commonly-used format is the *Newick format*. The name derives from the lobster restaurant in Dover (South Carolina), where it was developed (Felsenstein 2003).

The Newick format was originally designed for rooted trees. The description of the rooted tree starts at the tips and recursively proceeds through the internal nodes until the root. First, the tips are assigned labels. Then, two tips X and Y that are connected through a cherry are chosen, and their most recent common ancestor node is labelled by "(X : t_X, Y : t_Y)", where t_X (t_Y) denotes the length of the branch ancestral to node X (Y). Note that in the Newick format, "(X : t_X, Y : t_Y)" and "(Y : t_Y, X : t_X)" are equivalent. The tips X and Y together with their pendant branches are then deleted, meaning the node labelled with "(X : t_X, Y : t_Y)" becomes a tip. We then proceed recursively until we reach the root. In that way, each node is labelled with the Newick format of its descending subtree. The label of the root is the Newick format of the tree[3]. According to this definition, the Newick notation for the tree in Figure 6.4, middle, is "(((C : 1, D : 1) : 1, A : 2) : 1, B : 3)". Note that we may also write, for example, "(B : 3, ((C : 1, D : 1) : 1, A : 2) : 1)". In fact, we can swap expressions for subtrees to the immediate left and right of each comma, and thus, we can write our example tree in 2^3 different equivalent ways.

Unrooted trees can also be described in the Newick format by using an arbitrary internal node as the root node which would have three attached branches. Once the root node is reached, the three node labels are put together and separated by a comma; again, each ordering of the three node labels is allowed. Using these rules, the Newick format for the tree in Figure 6.4, left, is "((C : 1, D : 1) : 1, A : 2, B : 4)".

6.2.3 Counting trees

Often, we want to find the phylogenetic tree that best fits the available data. A naive tree reconstruction method may consider all possible trees to find the best tree. To assess the feasibility of this approach, we need to know how many different rooted (or unrooted) trees with n tips exist. We will start by counting the number of branches in a tree, which will facilitate counting the number of trees.

6.2.3.1 Counting branches

We can count the number of branches in an unrooted labelled tree with n tips by listing and counting all possible branches in that tree. Listing and counting, also called enumeration,

[3] If the root has an ancestral branch of length t_{root} attached, we extend the Newick format string by ": t_{root}".

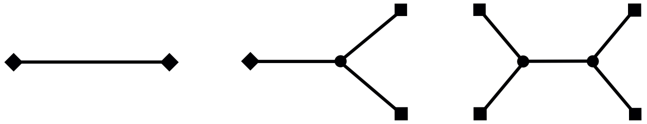


Figure 6.5: The unrooted trees with two, three, and four tips. The number of branches b_i for different tip numbers i is $b_2 = 1, b_3 = 3, b_4 = 5$.

is a handy approach for small trees. However, for large trees, this is computationally very inefficient, and we do not learn general patterns. For example, we cannot conclude if all trees on n tips have the same number of branches unless we count the branches for every single tree on n tips.

In what follows, we will derive an analytic formula for the number of branches, b_n, of a tree with n tips. Let us start with an example where $n = 2$. The two tips in an unrooted tree can only be connected with a single branch, thus $b_2 = 1$. For $n = 3$ tips, we have $b_3 = 3$ branches (see Figure 6.5). If we increase the number of tips by one, the number of branches increases by 2. This is because to add a tip, we need to break one of the existing branches into two and add an internal node to which the new tip with a new pendant branch attaches. Thus, for a tree with $n + 1$ tips we have $b_{n+1} = b_n + 2$ branches. In general, we have

$$b_n = b_2 + 2(n - 2) = 2n - 3 \tag{6.1}$$

for $n \geq 2$. In a rooted tree, we split one branch from the unrooted tree in two by adding the root node. This leads to

$$b_n^{root} = 2n - 3 + 1 = 2n - 2 \tag{6.2}$$

branches in the rooted tree for $n \geq 2$.

When we derived Equations (6.1) and (6.2), we employed the idea of a *proof by induction*. This technique is common in mathematics. With this technique, we can prove a formula that depends on an integer $n \in \mathbb{N}$. The proof consists of two steps. The first step establishes the formula for some small $m \in \mathbb{N}$, commonly $m = 1$ or $m = 2$. The second induction step then assumes that the formula holds for all $k < n$ and proves that it also holds for n using this assumption. With these two steps, it is proven that the formula holds for all $n \in \mathbb{N}$ with $n \geq m$. In the following example, we prove Equation (6.1) using proof by induction.

Theorem 6.2.1. *For an unrooted tree with n tips, the number of branches b_n is given by $b_n = 2n - 3$ (Equation (6.1)).*

Proof. We prove this statement using induction.

Hypothesis to prove: $b_n = 2n - 3$.

Base step: Check that the hypothesis holds for $n = 2$.

Yes, $b_2 = 1 = 2 \times 2 - 3$.

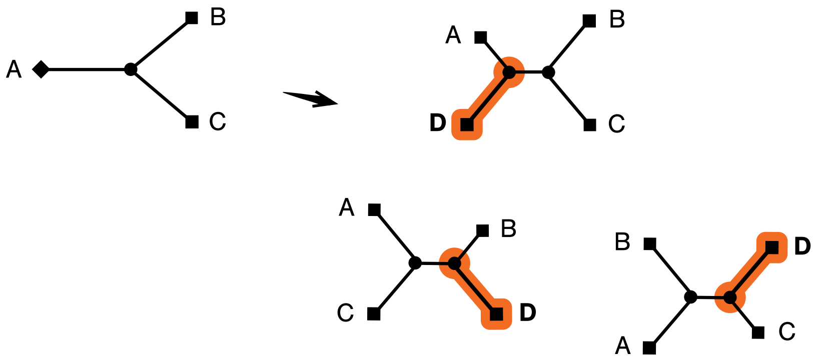


Figure 6.6: All unrooted four-tip trees that can be obtained from an unrooted labelled three-tip tree by attaching a new branch to any of the three existing branches.

Induction hypothesis: We suppose the formula holds for all $k < n$.
In particular, $b_{n-1} = 2(n-1) - 3$.

Inductive step: Given the induction hypothesis, show that the formula holds for n.
We have $b_n = b_{n-1}+2$, as explained above, adding a tip to a tree adds two new branches. Then, with the induction hypothesis, we establish $b_{n-1} + 2 = 2(n-1) - 3 + 2 = 2n - 3$.

Thus, our formula holds for all $n \geq 2$. □

6.2.3.2 Counting unrooted labelled trees

We now count the number of unrooted labelled trees τ_n on n tips with labels $l_1, \ldots, l_n$. For $n = 1$ tips in a tree, just one single unrooted tree is possible: $\tau_1 = 1$. For $n = 2$ and $n = 3$, there is also just one possible tree, so $\tau_2 = \tau_3 = 1$. To obtain any tree with $n = 4$ tips, we start with a tree with $n = 3$ tips with labels $l_1, \ldots, l_3$, and attach a new tip with label l_4 to any of the existing three branches (see Figure 6.6). Thus, $\tau_4 = 3$. Counting the number of trees with 5 tips, we get $\tau_5 = 15$. We now hypothesise that $\tau_n = 1 \times 3 \times 5 \times \ldots (2n - 5)$. The double factorial $m!!$ is defined as

$$m!! = \begin{cases} 1 \cdot 3 \cdot 5 \cdot \ldots \cdot (m-2) \cdot m & \text{for uneven } m \in \mathbb{N}, \\ 2 \cdot 4 \cdot 6 \cdot \ldots \cdot (m-2) \cdot m & \text{for even } m \in \mathbb{N}. \end{cases} \tag{6.3}$$

This means that our hypothesis is $\tau_n = (2n - 5)!!$, and we will prove this hypothesis by induction.

Theorem 6.2.2. *For a tree with n tips, the number of unrooted labelled trees τ_n is given by $\tau_n = (2n - 5)!!$.*

Proof.

Hypothesis to prove: $\tau_n = (2n-5)!!$ for $n \geq 3$.

Base step: Check that the hypothesis holds for $n = 3$.
Yes, there is only one labelled tree on 3 tips, thus $\tau_3 = 1$.

Induction hypothesis: We suppose the formula holds for all $k < n$.
In particular, $\tau_{n-1} = (2(n-1)-5)!!$.

Inductive step: Given the induction hypothesis, show that the formula holds for n.
Generally, note that each tree on n tips with labels $l_1, \ldots, l_n$ can be viewed as a subtree on $n-1$ tips with labels $l_1, \ldots, l_{n-1}$ plus the nth tip with label l_n attached to one of the branches of the $(n-1)$-tip tree. Importantly, starting with different $(n-1)$-tip trees or attaching the nth tip to different branches yields different trees on n tips. Trivially, if we start with two different trees on $n-1$ tips and attach the nth tip, we will always obtain different n-tip trees as the subtrees on $n-1$ tips are different. Second, if we start with the same $(n-1)$-tip subtree but attach the nth tip to different branches, we will get different trees as l_n will cluster with different tip labels. Thus, the number of n-tip trees τ_n is the number of $(n-1)$-tip trees times b_{n-1}, $\tau_n = \tau_{n-1} \times b_{n-1}$. Using the induction hypothesis together with the formula for b_n, we get $\tau_n = (2(n-1)-5)!! \times (2(n-1)-3) = (2n-5)!!$.

Thus, our formula holds for all $n \geq 3$. □

The number of labelled trees on n tips increases double-factorially with n, which roughly corresponds to exponential growth with $n \ln n$. Indeed, according to *Stirling's approximation*, for large n, we have $n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n = \sqrt{2\pi n} e^{n(\ln n - 1)}$. Furthermore, $(2n-5)!! = (2n-5)(2n-6)!! = (2n-5)2^{n-3}(n-3)!$, showing that the double factorial grows exponentially with $n \ln n$. Using the Landau notation (Box 7 on page 55), we can say that the number of trees τ_n is on the order $\mathcal{O}(e^{n \ln n})$. Table 6.1 shows some examples for the number of unrooted trees on n tips.

6.2.3.3 Counting rooted labelled trees

To count the number of rooted labelled trees on n tips, we note that each rooted labelled tree can be obtained from an unrooted labelled tree on n tips in which we choose one branch to be divided such that a root node is added. Importantly, choosing different unrooted labelled trees or different branches of these trees yields different rooted labelled trees. Thus, the number of rooted trees on n tips, τ_n^r, is

$$\tau_n^r = \tau_n \times b_n = (2n-5)!!(2n-3) = (2n-3)!!, \qquad (6.4)$$

which, again, is on the order $\mathcal{O}(e^{n \ln n})$.

Due to this large number of trees, finding the "best" tree from among all possible trees for a given dataset becomes very slow for large n. Therefore, we need to find smart ways to retrieve

Number of tips	Number of unrooted trees
4	3
5	15
6	105
7	945
8	10 395
9	135 135
10	2 027 025
11	34 459 425
⋮	⋮
20	221 643 095 476 699 771 875
⋮	⋮
50	10^{74}

Table 6.1: Number of unrooted trees on n tips.

the "best" tree in a reasonable time. The following section discusses common approaches for finding such trees.

6.3 Inferring phylogenies

In the early days of phylogenetics, observable phenotypic characteristics, or morphological traits, of the species were used to reconstruct phylogenetic trees. Species with similar morphological traits (e.g. the ability to fly) were thought to cluster together in the species tree, meaning that they were more closely related than species with very different morphological traits. However, this can lead to wrong phylogenies, for example, if *convergent evolution* of the phenotype occurred (evolution of the same morphological trait from different ancestors). A prominent example is flight. Not all animals that can fly are birds; there are some mammals and insects that can fly too, but are not closely related to birds and have evolved the ability to fly independently.

Nowadays, genetic sequence data largely replaces the morphological trait data when inferring phylogenies of living organisms (note though that for fossils, we still rely on morphology as we do not have genetic sequences, for an example, see Figure 10.10). Using genetic sequence data over morphological trait data has several advantages.

First, it is fairly straightforward to decide which data to include when considering genetic sequences: all nucleotides that are part of a multiple sequence alignment (MSA) of orthologues (or, more generally, homologues, when considering gene trees). On the other hand, using morphological traits relies on the choice of the trait and measurement.

Second, while molecular evolution models used for phylogenetic reconstruction still make simplifying assumptions, the modelling occurs where evolution happens, namely at the nucleotide, codon, and amino acid level, and each site contributes towards informing the tree reconstruction. In particular, the molecular evolution models typically assume neutral evolution and no selection (Section 5.3). If this does not hold, we can improve these probabilistic models of molecular evolution to account for selection, or we can use third codon position data with simple non-selection models as these positions may be assumed to evolve close to neutral (without selection) since such mutations rarely change the codon (Section 5.7). In contrast, morphology is where selection acts and where the consequences of molecular evolution become visible. Appropriate modelling of these selective processes and weighing the importance of the different characters when reconstructing phylogenies is far from trivial.

Another practical advantage of using sequences is that with the new high-throughput sequencing technologies, these data are much easier and cheaper to obtain than morphological trait data. For the latter, palaeontologists have to go on field trips to collect fossils and then take appropriate measurements. Thus, obtaining morphological data is the opposite of high-throughput.

Finally, when using genetic sequences, we can go beyond species and analyse pathogens or other individuals for which recording morphological traits is very hard.

Tree reconstruction methods generally take an MSA of homologous genetic sequences as an input. Alignment procedures have been described in detail in Chapter 3. When reconstructing phylogenies from MSAs, the rationale is to put "similar" sequences close in the tree (little evolution has occurred) and to place distant sequences very far apart in the tree (a lot of evolution has occurred). The word "similar" is put in quotation marks because there are several approaches to defining similarity when reconstructing phylogenetic trees. In this book, we will introduce the three main approaches. We will discuss tree reconstruction methods designed for each approach, as well as the advantages and disadvantages of each approach.

Phenetic approaches: (distance-based methods) infer the tree based on pairwise similarity of genetic sequences. The pairwise distance between two sequences can be derived under a molecular evolution model (Section 5.4).

Cladistic approaches: (parsimony methods) group organisms based on how many shared characteristics they have without relying on an explicit molecular evolution model (but see Section 6.3.3.3).

Probabilistic approaches: (maximum likelihood and Bayesian methods) assume an explicit probabilistic model of evolution for the underlying data and group the organisms based on the likelihood. The underlying transition probability matrices of the molecular evolution models are introduced in Chapter 5.

sequence 1, s_1: `TCACACCT`
sequence 2, s_2: `ACAGACTT`
sequence 3, s_3: `AAAGACTT`
sequence 4, s_4: `ACACACCC`

Table 6.2: Toy MSA on which the tree reconstruction methods are illustrated.

H	s_1	s_2	s_3	s_4
s_1	–	3	4	2
s_2		–	1	3
s_3			–	4
s_4				–

Table 6.3: Hamming distance matrix for the MSA in Table 6.2.

We will illustrate the ideas behind the different tree reconstruction approaches and methods using the MSA in Table 6.2 for sequences taken from four individuals.

6.3.1 Phenetic approach: Distance-based methods

The idea of distance-based methods is to cluster sequences that are most similar to each other. Similarity is measured as the pairwise sequence distance. As described in Chapter 5, we can choose between different pairwise distance measures, such as, for example, the Hamming distance, JC69 distance or HKY distance (see Table 5.2). Distances can also be defined and calculated based on morphological characters, and the distance-based methods can be applied to such distance measures equivalently.

When reconstructing the tree, we first calculate the distance between each pair of sequences. Then, the phylogeny is reconstructed such that a pair of sequences with a small distance are close to each other in the tree. A general drawback of distance-based methods is that they only use pairwise sequence distances, and no higher-order relationships (shared by more than two sequences in the sample) are considered.

Here, we demonstrate the phenetic tree reconstruction using the Hamming distance measure applied to the MSA shown in Table 6.2. The Hamming distance between each pair of sequences is put into a matrix called the *distance matrix* (Table 6.3). We put the distance $d(s_i, s_j)$ $(i < j)$ between sequence s_i and s_j in matrix position (s_i, s_j).

The matrix with all pairwise distances is symmetric; the distance between s_1 and s_2 is the same as the distance between s_2 and s_1. Therefore, we only fill out the upper triangle in the distance matrix. Furthermore, note that the distance of a sequence to itself, say s_1 to s_1, is 0.

JC69	s_1	s_2	s_3	s_4
s_1	–	0.52	0.82	0.30
s_2		–	0.14	0.52
s_3			–	0.82
s_4				–

Table 6.4: JC69 distance matrix for the MSA in Table 6.2.

In the distance matrix, we put a – on the diagonal. We label this matrix with an "H" in the upper left corner to indicate that all pairwise distances were calculated using the Hamming distance measure.

Ideally, we want the distance between two sequences to represent the amount of evolution that has occurred between them. One way to quantify this is the genetic distance (Section 5.4): the expected number of substitutions a single site underwent in the time separating the sequences. This is different from the Hamming distance, which only counts observed differences. As derived in Section 5.4.2, the pairwise distance formula under JC69 is $\hat{d} = -\frac{3}{4}\log(1 - \frac{4}{3}\hat{p})$ where $\hat{p} = {}^{x}/_{n}$ and x is the number of differences between two sequences (the Hamming distance), and n is the sequence length. The distance matrix under the JC69 model for the toy MSA in Table 6.2 is shown in Table 6.4.

There are two different classes of distance-based methods to infer a tree based on the distance matrix. The first class is referred to as *algorithmic methods*. These methods cluster together sequences that are separated by the smallest distance according to the distance matrix in a greedy way, meaning that in each step, they cluster according to the best choice at that moment, with the idea that many sequential best local choices should lead to a good global choice. This means the sequences separated by the smallest pairwise distances are picked sequentially and clustered in the tree. Such methods are very fast and are often used if the distance matrices are large. Examples of algorithmic methods are UPGMA (Sokal and Michener 1958) or neighbour-joining algorithms (Saitou and Nei 1987).

The second class of methods is referred to as *optimality methods*. These methods try to minimise the difference between the sequence distance matrix and the inferred tree distance matrix (Fitch and Margoliash 1967; Cavalli-Sforza and Edwards 1967). The distances in the tree distance matrix are the sum of branch lengths on the paths between each pair of tips. These methods can be very slow as they typically have to consider all possible trees to find the tree that minimises the difference. We will now introduce the two classes of methods in more detail.

6.3.1.1 Algorithmic approach: UPGMA method

A classic distance-based (algorithmic) method is the *UPGMA (Unweighted Pair Group Method using Arithmetic means)* algorithm (Sokal and Michener 1958). The UPGMA algorithm constructs ultrametric trees, meaning it is only suited for data sampled at one point in time. UPGMA makes the implicit assumption that the (genetic) data evolved according to a *strict molecular clock*, that is, the substitution rates were the same for all branches in the tree at all times. Then, the genetic distances in the distance matrix map onto an ultrametric tree where branch lengths are proportional to calendar time.

If the assumption of a strict molecular clock is violated or sequences are sampled at different time points, then methods inferring non-ultrametric trees are required. An algorithmic distance-based method reconstructing non-ultrametric trees is the *neighbour-joining algorithm* (Saitou and Nei 1987). This algorithm infers unrooted trees with branch lengths in units of numbers of substitutions without proportionality to calendar time.

The different algorithmic approaches have in common that iteratively, pairs of nodes are joined together to obtain a tree. Since the UPGMA method provides an easy and intuitive understanding of algorithmic distance-based methods, we will present this method below. In particular, we provide a step-by-step description of the UPGMA algorithm, see Algorithm 1. Such a description can serve as a blueprint for implementing the algorithm in code and is called pseudocode.

Example of UPGMA with the Hamming distance matrix Here, we show how the UPGMA algorithm (Algorithm 1) works for our example Hamming distance matrix, shown in Table 6.3.

Iteration 1 In step 1 of the UPGMA algorithm, we look up the minimal pairwise distance in the Hamming distance matrix. In our example distance matrix, the minimal distance is $d(s_2, s_3) = 1$, the distance between nodes s_2 and s_3.

In step 2 of the UPGMA algorithm, we join these two nodes and introduce a new node $s_{2,3}$. This means that the currently reconstructed tree is a cherry plus the unconnected nodes as shown in Figure 6.7. The distance from the new node $s_{2,3}$ to the tips s_2 and s_3 is ${}^{d(s_2,s_3)}/_2 = 0.5$. We further set $n_{2,3} = 2$.

In step 3, we calculate the distances from the new node to the remaining nodes. For example, to calculate the distance from the new node $s_{2,3}$ to the node s_1, we calculate $d(s_1, s_{2,3}) = \frac{n_2 d(s_2,s_1)+n_3 d(s_3,s_1)}{n_2+n_3} = \frac{1\times3+1\times4}{1+1} = \frac{7}{2} = 3.5$. The distances between all remaining nodes (in this case, the distance between s_1 and s_4) remain the same. The new distance matrix obtained from step 4 is shown in Table 6.5.

```
Input: Distance matrix for $n$ sequences. We refer to each sequence $s_1, \dots, s_n$ as a node.
Output: A rooted ultrametric phylogenetic tree.
begin
    initialise the size of each node $s_i$ as $n_i = 1$;
    initialise the tree as the set of unconnected nodes $s_i, i = 1, \dots, n$;
    while the distance matrix includes at least 2 nodes do
        (step 1) choose nodes $s_i$ and $s_j$ such that $d(s_i, s_j)$ is the smallest entry in the
          distance matrix (in case of several minima, choose one uniformly at random);
        (step 2) join nodes $s_i$ and $s_j$ of the current tree to form a new node $s_{i,j}$ with size
          $n_{i,j} = n_i + n_j$; set the branch length between $s_i, s_{i,j}$ and between $s_j, s_{i,j}$ such
          that all tips descending from $s_{i,j}$ have the same distance $d(s_i,s_j)/2$ to $s_{i,j}$;
        if the distance matrix includes only 2 nodes then
            return the tree with the root $s_{i,j}$ as the result and finish.
        end
        (step 3) include node $s_{i,j}$ into the distance matrix, with
          $d(s_m, s_{i,j}) = \frac{n_i d(s_i,s_m) + n_j d(s_j,s_m)}{n_i + n_j}$, where $s_m$ is a node in the distance matrix;
          delete nodes $s_i$ and $s_j$ from the distance matrix;
    end
end
```

Algorithm 1: The UPGMA algorithm.

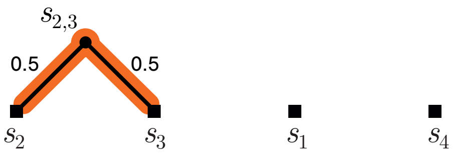


Figure 6.7: Intermediate UPGMA tree after one iteration.

H	s_1	s_4	$s_{2,3}$
s_1	–	2	3.5
s_4		–	3.5
$s_{2,3}$			–

Table 6.5: Intermediate UPGMA distance matrix after one iteration.

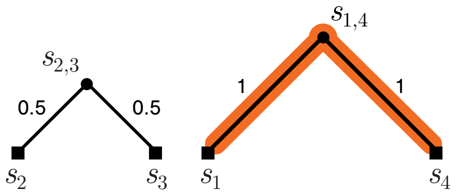


Figure 6.8: Intermediate UPGMA tree after two iterations.

H	$s_{2,3}$	$s_{1,4}$
$s_{2,3}$	–	3.5
$s_{1,4}$		–

Table 6.6: Intermediate UPGMA distance matrix after two iterations.

Iteration 2 The minimal distance between a pair of nodes in the new matrix is between s_1 and s_4, $d(s_1, s_4) = 2$. We create a new node $s_{1,4}$ and set $n_{1,4} = 2$. The intermediate UPGMA tree is shown in Figure 6.8, and the new distance matrix is shown in Table 6.6.

Iteration 3 Now, the minimal distance is between $s_{2,3}$ and $s_{1,4}$, $d(s_{2,3}, s_{1,4}) = 3.5$. In step 2 of the algorithm, we create the new node $s_{2,3,1,4}$, which is the root of the tree, with distance ${}^{3.5}/_{2} = 1.75$ to all tips. As only two nodes $s_{2,3}, s_{1,4}$ are in the distance matrix, the algorithm terminates, outputting the tree shown in Figure 6.9.

Properties of the UPGMA trees The UPGMA tree distance matrix (the distance matrix composed of the sum of branch lengths on the path between each pair of tips in the UPGMA tree) is shown in Table 6.7.

This tree distance matrix differs slightly from the original sequence distance matrix (Table 6.3). The reason is that pairwise distances in the sequence distance matrix do not generally correspond to an ultrametric tree. The UPGMA (and other distance-based methods) construct a tree representing the distances in the matrix as well as possible. The UPGMA has the desired property that if the input distance matrix equals the tree distance matrix of an ultrametric tree, then precisely this ultrametric tree will be returned by UPGMA. We will prove this property of UPGMA in the following theorem.

Theorem 6.3.1. *Let D_N be a distance matrix of dimension $N \times N$. Assume there exists an ultrametric tree T_N with a tree distance matrix that is equal to D_N. Then, the UPGMA algorithm, as defined in Section 6.3.1.1 with input matrix D_N, will return T_N.*

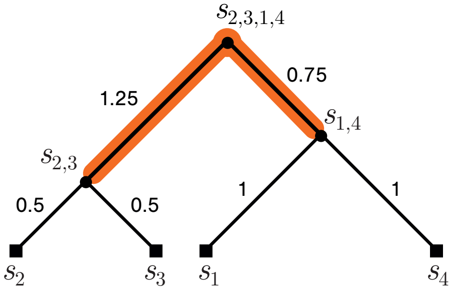


Figure 6.9: Final UPGMA tree.

	s_1	s_2	s_3	s_4
s_1	–	3.5	3.5	2
s_2		–	1	3.5
s_3			–	3.5
s_4				–

Table 6.7: Tree distance matrix for the reconstructed UPGMA tree.

Proof. We prove this statement using induction as described in Section 6.2.3.1.

Hypothesis to prove: If the distance matrix D_N equals the tree distance matrix of an ultrametric tree T_N, the UPGMA algorithm with input $\mathrm{D}_N^{n_1,\ldots,n_N}$, where $n_1, \ldots, n_N$ are arbitrarily chosen integers (rather than 1 as in the classic UPGMA presented above), returns the tree T_N.

Base step: Let $N = 2$, we consider an ultrametric tree T_2 with two branches of length $^{d(s_1,s_2)}/_2$. The distance matrix D_2 has only one entry $d(s_1, s_2)$, which is also the smallest entry. The UPGMA algorithm (step 2) results in an ultrametric tree with two tips and branch lengths $^{d(s_1,s_2)}/_2$, obtaining the tree T_2. In particular, this does not depend on the values of n_1, n_2. Thus, the hypothesis holds for $N = 2$.

Induction hypothesis: Suppose the hypothesis holds for all $k < N$.

Inductive step: Given the hypothesis holds for $k < N$, we need to prove the hypothesis for N. According to step 1 of the UPGMA algorithm, we choose the two nodes s_i and s_j in D_N whose entry $d(s_i, s_j)$ in the distance matrix is the smallest. This node is joined to a cherry C with branch lengths $^{d(s_i,s_j)}/_2$. Since D_N is the tree distance matrix for T_N, exactly this cherry also appears in T_N. Further, since D_N is a tree distance matrix for

Box 27: Statistical consistency

An estimation method is *statistically consistent* if the true parameter is returned when the method is presented with an infinite amount of data. Put more formally, a method is statistically consistent if, for all $\varepsilon > 0$, we have

$$\lim_{n\to\infty} P(||\hat{T}_n - T|| < \varepsilon) = 1, \tag{B27.1}$$

where n is the number of data points, T is the true parameter, and $\hat{T}$ is the inferred parameter. $||\cdot||$ is a metric measuring how different the inferred parameter is from the true parameter. This property is desirable for inference methods because they will return the correct answer provided enough data.

In our context, n is the sequence length, T is the true tree on which the sequence data were generated, and $\hat{T}$ is the inferred tree. This means that if our tree inference method is statistically consistent, the probability that we obtain a tree that differs less than ε from the true tree tends to 1 when the input sequence length tends to infinity.

an ultrametric tree, we know that

$$d(s_i, s_m) = d(s_j, s_m) \tag{6.5}$$

is true for all nodes s_m, $m \neq i, j$. According to step 3 in the UPGMA algorithm, the new node $s_{i,j}$ has the following distance to the remaining nodes s_m:

$$d(s_m, s_{i,j}) = \frac{n_i d(s_i, s_m) + n_j d(s_j, s_m)}{n_i + n_j} \overset{(6.5)}{=} \frac{(n_i + n_j) d(s_i, s_m)}{n_i + n_j} = d(s_i, s_m). \tag{6.6}$$

Thus, the new matrix D_{N-1} is a distance matrix for the ultrametric tree T_N without leaf j, T_{N-1}. According to the induction hypothesis, UPGMA with input D_{N-1} returns T_{N-1}. Thus, the UPGMA on D_N returns the tree T_{N-1} with tip i replaced by cherry C, thus, it returns T_N.

Thus, the hypothesis holds for all $N \geq 2$. □

Furthermore, UPGMA is statistically consistent (see Box 27 on page 152) for ultrametric trees, which means when the method is presented with an infinite amount of data (infinitely long sequences), the returned tree is the one on which the data were actually generated (Felsenstein 2003; Gascuel and McKenzie 2004).

Runtime of the UPGMA method The advantage of algorithmic distance-based methods is their speed. To reconstruct a tree with n tips, we need to perform $n - 1$ iterations of the algorithm (in each iteration, two nodes join into one). Within each iteration, a distance matrix has to be set up and searched, which has a time complexity on the order of $\mathcal{O}(n^2)$. This means that a naive approach would need to perform on the order of $\mathcal{O}(n^3)$ calculations to obtain the tree. However, a clever implementation can achieve faster runtimes (see, e.g. Murtagh

(1984)). The phylogenetic reconstruction methods presented in later sections will be much slower than the polynomial-time algorithmic distance-based methods; they have at least exponential runtime, $\mathcal{O}(e^n)$, since all trees on n tips need to be checked for optimality. For many tens of thousands of sequences, only algorithmic distance-based methods or approximations to the methods with exponential runtime will be fast enough to infer a tree in a reasonable time.

6.3.1.2 Optimality approach: a least squares method

The *least squares method* searches for a tree that minimises the squared difference between the sequence distance matrix and the tree distance matrix (Fitch and Margoliash 1967; Cavalli-Sforza and Edwards 1967). In other words, the algorithm minimises the scoring function

$$S = \sum_{i=1}^{n} \sum_{j=i+1}^{n} w_{i,j}(D_{i,j} - d_{i,j})^2, \tag{6.7}$$

where D is the sequence distance matrix, d is the tree distance matrix for the proposed tree, and $w_{i,j}$ are the weights ($w_{i,j}$ may be, for example, 1 or $^1/_{D_{i,j}}$).

The least squares method has the desirable property of being statistically consistent, as shown in the following theorem.

Theorem 6.3.2. *Least squares methods with distances between sequences obtained using the maximum-likelihood estimator (Section 5.4.3) and assuming the model under which the sequences evolved are statistically consistent.*

Proof. Consider a fixed tree with branch lengths measured in units of substitutions, meaning that one substitution is expected to happen within one time unit. Let sequences evolve on this tree under some model M. Let maximum likelihood distances be calculated using the model M. A maximum likelihood estimator is statistically consistent (Newey and McFadden 1994). Here, that means with increasing sequence length, the sequence-induced distance matrix $\hat{D}_n$ approaches the tree-induced distance matrix D, $\lim_{n\to\infty} P(||\hat{D}_n - D|| < \varepsilon) = 1$.

The mapping between a tree T and its distance matrix D is a bijection. We define the tree metric as $||\hat{T}_n - T|| = ||\hat{D}_n - D||_2 = \sum_{i=1}^{n}\sum_{j=i+1}^{n}(\hat{D}_{i,j} - D_{i,j})^2$, where the latter is the function to be minimised in the least squares method. Then, $\lim_{n\to\infty} P(||\hat{T}_n - T|| < \epsilon) = \lim_{n\to\infty} P(||\hat{D}_n - D||_2 < \varepsilon) = 1$. This establishes that the least squares method is statistically consistent. □

Least squares algorithms are much slower than the algorithmic approaches presented before. They search the whole space of trees to find the tree that minimises the criterion, meaning the runtime is $\mathcal{O}(e^{n \ln n})$. Furthermore, the least squares optimisation problem is an NP*-hard* problem (Day 1987), see Box 28 on page 154.

Box 28: *NP*-completeness and *NP*-hardness

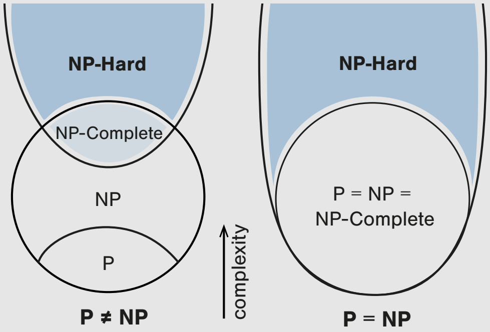
A *decision problem* is a question that can be posed as a yes-no question, for example, "Is there a tree inducing a distance matrix with a least squares difference to a given sequence distance matrix of less than x?". The corresponding *optimisation problem* would be "What is the tree with the minimal least squares difference to the given sequence distance matrix?"

In computational complexity theory, P stands for polynomial computation time. The set of decision problems P is the set of decision problems that can be solved in polynomial time with respect to the dataset size. This means that the number of computations (time) required to solve a problem in P with input size n is on the order of $\mathcal{O}(n^k)$, where k is some fixed input-independent number (e.g. $k = 3$ for the UPGMA above).

NP stands for nondeterministic polynomial time. The set NP is the set of decision problems for which it is possible to verify whether a particular proposal is a solution in polynomial time. In the least squares method, it is easy to determine if, for a given tree, the least squares difference of its distance matrix to the original sequence distance matrix is less than some threshold x. Thus, this decision problem is in NP.

By definition, $P \subseteq NP$ (P is a subset of NP), but whether $P = NP$, whether decision problems for which a solution can be verified in polynomial time can also be solved in polynomial time, is currently unknown, and one of the major conundrums in computer science.

A decision problem X is NP*-complete* if an algorithm solving X could also be used to solve all other problems in NP, potentially employing a polynomial time transformation of the algorithm. As a consequence of the definition of NP-completeness, if one NP-complete problem can be solved in polynomial time, then all problems in NP can be solved in polynomial time (and thus $P = NP$). Proving or disproving $P = NP$ is one of the 7 Millennium Prize Problems announced in 2000 (Jaffe 2006). The first person to solve it will be awarded 1 million US dollars. A Venn diagram can nicely display the connections between P, NP, and NP-complete:



A popular example of an NP-complete problem is the *travelling salesman problem*. The travelling salesman problem considers k cities (e.g. capitals of Europe) that a salesman has to visit. In the travelling salesman decision problem, we want to know whether the salesman can visit all the cities on a path shorter than length L.

A problem H is NP*-hard* if an algorithm solving it can also solve a NP-complete problem X, possibly employing a polynomial time transformation to adopt the algorithm for H such that it solves X. Thus, any NP-complete problem is also NP-hard. However, an NP-hard problem does not need to be in the class NP. In particular, a solution may not be verifiable in polynomial time, or the problem may not be a decision problem. In fact, an NP-hard problem may be an optimisation problem. An example is the optimisation version of the travelling salesman problem, in which we want to know the shortest path for a salesman to visit all k cities. Given that a decision problem is NP-complete, the corresponding optimisation problem is NP-hard: we can answer the decision problem (whether a solution $\leq L$ exists) using an algorithm that solves the optimisation problem.

6.3.2 Cladistic approach: Parsimony method

We now introduce the cladistic approach, going back to Edwards and Cavalli-Sforza (1964). It groups sequences based on how many characteristics they share. While phenetic methods group sequences based on pairwise similarity, ignoring relationships of more than two sequences at a time, the cladistic method accounts for higher-order sequence relationships.

Using the cladistic approach in tree inference leads to an unrooted tree, the *maximum parsimony tree*. For a given MSA, the maximum parsimony tree is an unrooted tree on n tips with the lowest *parsimony score* among all unrooted trees on n tips. The parsimony score of a tree is defined as the minimal number of changes (such as nucleotide substitutions for a nucleotide MSA) required to explain the MSA on the tree. Thus, the cladistic approach aims to determine the tree requiring the lowest number of changes.

6.3.2.1 Parsimony score example

We illustrate the concept of parsimony with our example MSA. First, since our MSA contains five polymorphic sites (shown in orange below), we require at least five substitutions, and 5 is the lower bound for the parsimony score.

sequence 1: TCACACCT
sequence 2: ACAGACTT
sequence 3: AAAGACTT
sequence 4: ACACACCC

We will show how to calculate the parsimony score for the UPGMA tree (Figure 6.10) reconstructed earlier. Note that this UPGMA tree is rooted. In fact, the parsimony score is calculated on rooted trees. However, as we explain below, the parsimony score for all rooted trees with the same underlying unrooted tree is the same, meaning that the parsimony method cannot distinguish among these trees.

We start by looking at each site and determining how many substitutions are required to achieve the configuration at the tips. For example, sequences s_2 and s_3 have an A at the first site, so no substitution is required on the branches leading from the tips to their common ancestor. On the other hand, s_1 has a T, but s_4 has an A at the first position. Thus, one substitution is required; there may have been a substitution from T to A on the branch leading to s_4 or a substitution from A to T on the branch leading to s_1. To decide, we look at the other sequences. Since s_2 and s_3 both have an A, it is more parsimonious that the sequence s_1 changed rather than all the other sequences. We thus assign an A $\rightarrow$ T substitution to the branch between the s_1 tip and the common ancestor of s_1 and s_4. The parsimony score for the first site is 1 since we need one substitution to explain the MSA at this site. We proceed in the same way for all the remaining sites, assigning and counting substitutions. The sum of the parsimony scores for each site is the parsimony score for the tree, which in this case is

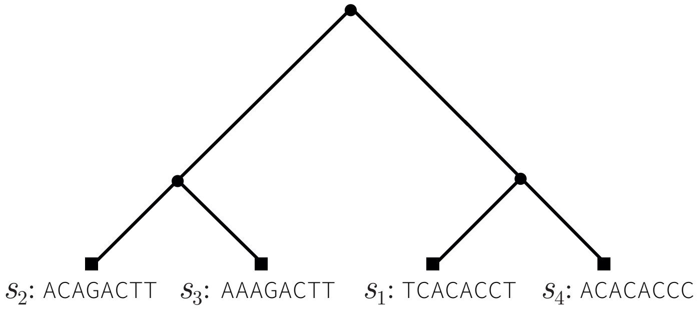


Figure 6.10: Tree obtained by the UPGMA algorithm with the sequences at the tips. This tree will be used to illustrate the parsimony score.

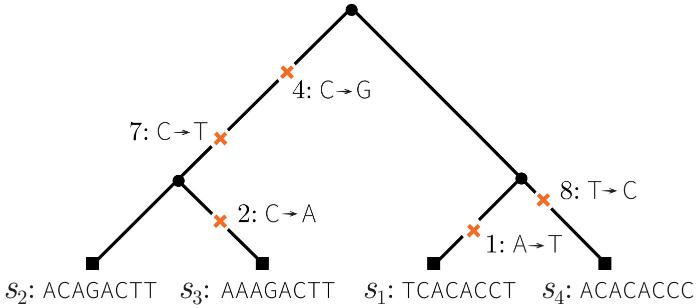


Figure 6.11: Tree obtained by the UPGMA algorithm with the sequences at the tips and a minimal number of required substitutions assigned to the branches.

five (Figure 6.11). Note that here, the optimal assignment was not obtained systematically; we merely illustrated the concept. In Section 6.3.2.3, we will present a fast and rigorous algorithm to obtain the parsimony score for a site.

6.3.2.2 Rooted versus unrooted trees

No matter where we root the tree, we obtain the same parsimony score through the same substitutions. In other words, the parsimony score will stay the same if we omit the root in the UPGMA tree and then reroot by choosing another branch in the unrooted tree. This is because we can use the same substitutions as in the original tree to explain the sequences in the rerooted tree. In Figure 6.12, we display all possibilities for rerooting the UPGMA tree; all these trees have the same unrooted tree.

Our example MSA has five polymorphic sites; thus, the maximum parsimony tree has a score of at least 5. Since the UPGMA tree has a parsimony score of 5, we know that the corresponding unrooted tree is one of the maximum parsimony trees. Thus, we found a maximum

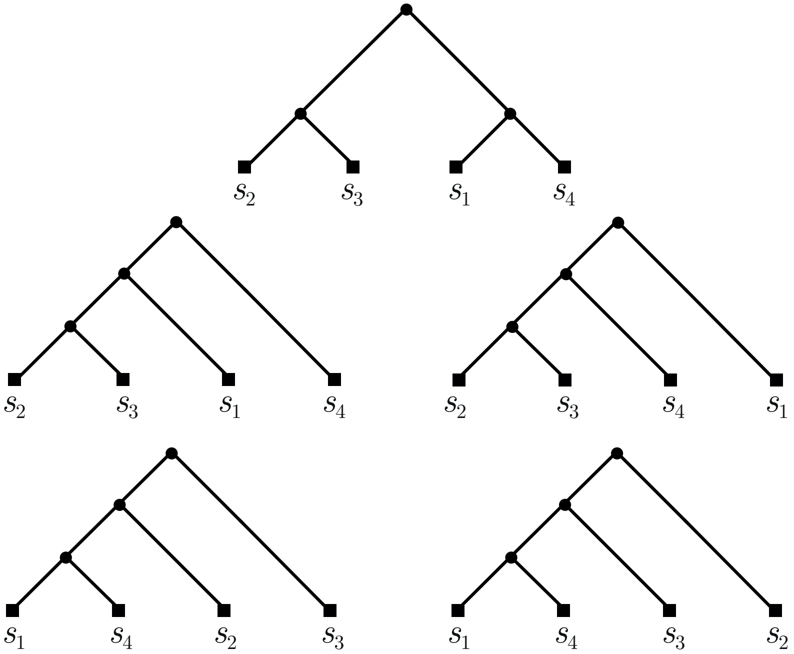


Figure 6.12: Tree obtained by the UPGMA algorithm and all possible rerooting of this tree. All five rooted trees have the same underlying unrooted tree and, therefore, the same parsimony score.

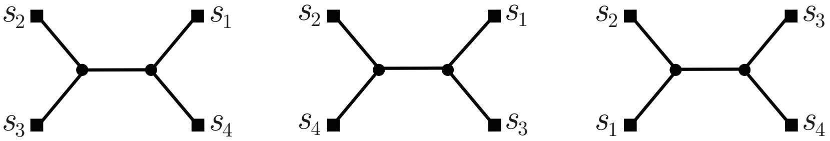


Figure 6.13: The three unrooted labelled trees on four tips.

parsimony tree by just looking at one rooted tree (though there may be other trees with the same score). Often, a maximum parsimony tree has a score greater than the number of polymorphic sites in the MSA. This means that one has to consider the parsimony score of all unrooted trees to determine the maximum parsimony tree. Table 6.1 states the number of unrooted trees on n tips and Figure 6.13 shows all unrooted trees on four tips.

6.3.2.3 Fitch algorithm

To determine the parsimony score of a tree, we can assign all possible ancestral sequences to the internal nodes and then count the number of substitutions for each assignment of ancestral sequences. For n nucleotide sequences of length m, we have a rooted tree on n tips and $n-1$ internal nodes, thus 4^{n-1} possible nucleotide assignments per site. Given a site assignment, we need to determine if the nucleotide changed or not for each of the $2n-2$ branches, meaning we have to perform $2n-2$ operations. Thus, overall, we need to perform $m \cdot 4^{n-1}(2n-2)$ operations to obtain the parsimony score (see also Section 6.3.2.1). Computing the parsimony score for all these assignments would be very slow. Such an approach trying all possibilities is called *brute-force*.

To calculate the parsimony score for a given tree, we can instead use a smarter approach involving dynamic programming. We already encountered dynamic programming for pairwise sequence alignment in Chapter 3. Recall that the idea behind the dynamic programming approach is to break down the problem into a collection of smaller subproblems, solve the small subproblems first, store the results, and then use them to solve the bigger problem. This strategy is superior to brute-force approaches, which essentially evaluate the same subproblem multiple times. In a phylogenetic context, dynamic programming translates to recursively solving the subproblem for a subtree and combining the results on the subtrees to obtain the result for the full tree.

A fast dynamic programming algorithm for computing the parsimony score is the *Fitch algorithm* (Fitch 1971), the main idea of which is to recursively calculate the parsimony score on subtrees and then combine the subtree results to obtain the parsimony score for the full tree. In Algorithm 2, we outline the Fitch algorithm in pseudocode form for nucleotides. It works analogously for amino acids or codons.

Note that in this algorithm, we denote $\cap$ as taking the intersection of two sets, that is, all elements in both sets. $\cup$ denotes the union, that is, all elements that belong to at least one of the sets. $\emptyset$ is the empty set. For two disjoint sets A and T, $\{A\} \cup \{T\} = \{A, T\}$; $\{A\} \cap \{T\} = \emptyset$.

We will now prove by induction over the number of tips n in the tree that the Fitch algorithm outputs the parsimony score. We provide proof for one site, $m = 1$. This directly completes the proof for $m > 1$ since the parsimony score for many sites is the sum of parsimony scores for each of them.

Theorem 6.3.3. *For any n and $m = 1$, the Fitch algorithm outputs the parsimony score S and all optimal root nucleotides. An optimal root nucleotide is a nucleotide that explains the tip nucleotides with S substitutions.*

Proof.

Hypothesis to prove: For any n and $m = 1$, the Fitch algorithm outputs the parsimony score S and all optimal root nucleotides.

```
Input: An unrooted phylogenetic tree and an MSA of $n$ sequences of length $m$,
    corresponding to the $n$ tips of the tree.
Output: Parsimony score of the tree, the minimal number of substitutions required to
    explain the sequences at the tips.
begin
    root the tree at an arbitrary branch;
    $k \leftarrow 0$;
    while the root has no sequence assigned do
        choose a node in the tree where all the descending nodes (all nodes on a path
          down to a tip) have sequences assigned;
        assign a sequence to the chosen node in the following way:
        for $i = 1, \ldots, m$ do
            let $C_l$ and $C_r$ be the sets of nucleotides assigned to the two direct descendants
              of the chosen node for site $i$;
            if $C_l \cap C_r \neq \emptyset$ then
                assign $C_l \cap C_r$ to site $i$ of the chosen node and keep $k$ unchanged;
            else
                assign $C_l \cup C_r$ to site $i$ of the chosen node and set $k \leftarrow k + 1$;
            end
        end
    end
    return $k$ as the parsimony score.
end
```

Algorithm 2: Fitch algorithm to compute the parsimony score of a tree.

Base step: Check that the hypothesis holds for $n = 2$.
If the nucleotides differ, the parsimony score is 1, and an optimal root nucleotide is either of the tip nucleotides. If the nucleotides are the same, the parsimony score is 0, and the optimal root nucleotide equals the tip nucleotide. The Fitch algorithm returns precisely these values. This completes the base step.

Induction hypothesis: Suppose that the Fitch algorithm returns the parsimony score and all optimal root nucleotides for all trees with k tips, $k < n$.

Inductive step: In the inductive step, we now show that the Fitch algorithm returns the parsimony score and all optimal root nucleotides for all trees with n tips. We split the rooted tree on n tips into two rooted subtrees, 1 and 2, by deleting the root and the two adjacent branches. By applying the induction hypothesis, we obtain the parsimony score S_1 and S_2 and the optimal root nucleotides using the Fitch algorithm for the two subtrees. The parsimony score of the n-tip tree is at least $S_1 + S_2$.

Case 1: The intersection of the sets of optimal root nucleotides for both subtrees is non-empty. This means that any nucleotide of the intersection can serve as a root nucleotide, which leads to $S_1 + S_2$ changes. Thus, the parsimony score attains the lower bound $S_1 + S_2$, and all nucleotides in the intersection belong to the set of optimal nucleotides. Next, we show that a nucleotide X not in the intersection cannot be an optimal root nucleotide:

(i) if both subtrees are assigned one of their optimal root nucleotides, then X has to change on at least one branch adjacent to the root, and the number of substitutions is at least $S_1 + S_2 + 1$;

(ii) if one subtree has a non-optimal root nucleotide assigned, then the number of substitutions required in that subtree is $S_1 + 1$, and thus, the overall number of substitutions is at least $S_1 + S_2 + 1$, which is bigger than the parsimony score of $S_1 + S_2$.

Case 2: The intersection of the optimal root nucleotides for both subtrees is empty. We show that in this case, the parsimony score is $S_1 + S_2 + 1$, and the set of nucleotides in the union is the set of optimal root nucleotides.

We first show that the parsimony score is $S_1 + S_2 + 1$, which can be obtained with a root nucleotide from the union of optimal root nucleotides of the subtrees. We distinguish three cases:

(i) if the two subtrees each have any one of their optimal root nucleotides assigned, the subtrees contribute $S_1 + S_2$ substitutions, and joining the two subtrees requires an additional substitution — where the root nucleotide is one of the two subtree root nucleotides, — leading to $S_1 + S_2 + 1$ substitutions;

(ii) if one subtree is assigned a non-optimal root nucleotide X, then it contributes at least $S_1 + 1$ substitutions and we get an additional S_2 substitutions from the other subtree, meaning that at least $S_1 + S_2 + 1$ substitutions are required;

(iii) if both subtrees are assigned non-optimal root nucleotides, they contribute $S_1 + S_2 + 2$ substitutions, which is bigger than the parsimony score.

Thus, $S_1 + S_2 + 1$ is the parsimony score. The first case shows that this score can be obtained with the root nucleotide from the union.

It remains to be shown that the union of the optimal subtree root nucleotides is the set of all optimal root nucleotides. Suppose nucleotide X is not in the union, then

(i) if both subtrees are assigned one of their optimal root nucleotides, then X has to change on both branches adjacent to the root, and the number of substitutions is $S_1 + S_2 + 2$;

(ii) if one subtree has a non-optimal root nucleotide assigned, then the number of substitutions required in that subtree is at least S_1+1. Now consider the two cases for the other subtree:

(a) if the other subtree has an optimal nucleotide assigned, it requires one substitution from X to that optimal nucleotide plus S_2 substitutions within that subtree, leading to at least $S_1 + S_2 + 2$ substitutions;

(b) the other subtree has a non-optimal nucleotide assigned, leading to at least S_2+1 substitutions in that subtree, and overall leading to at least S_1+S_2+2 substitutions.

Thus, X is not an optimal root nucleotide.

□

The Fitch algorithm is very fast as it traverses each of the $n-1$ internal nodes once, assigning a state to each site at each node, resulting in a runtime of the order $\mathcal{O}(nm)$ where n is the number of sequences and m is the sequence length (in our example $n = 4, m = 8$). Thus, the parsimony score on a given phylogenetic tree is calculated in linear time in n.

6.3.2.4 Example of the Fitch algorithm

We continue with our example alignment from above. We consider the three possible unrooted trees on four tips, shown in Figure 6.13. Following the Fitch algorithm, we now assign sequences to internal nodes (Figure 6.14). When a set of nucleotides, say A and G, is assigned to a particular node at a site, we write $\{\mathtt{A},\mathtt{G}\}$. For the middle and bottom trees, seven substitutions are required to explain the sequences at the tips; for the top tree, only five substitutions are required, meaning that the top tree is the only maximum parsimony tree.

6.3.2.5 Time complexity of the parsimony method

The Fitch algorithm calculates the parsimony score for a single tree, the input tree. However, we have to calculate the parsimony score for each possible unrooted tree on n tips. For $n = 4$, this is easy, as only three trees have to be considered, but in general, the number of trees (see Section 6.2.3.2) — and as a consequence, the runtime, — increase drastically with n. One may thus ask if there is a fast way to compute the maximum parsimony tree without considering each unrooted tree. The answer is no unless $P = NP$: we can show that finding the maximum parsimony tree is an NP-hard problem (Foulds and Graham 1982), meaning we have to essentially calculate the parsimony score for all possible trees on n tips. Among all possible trees, the trees with the lowest parsimony score are the maximum parsimony trees.

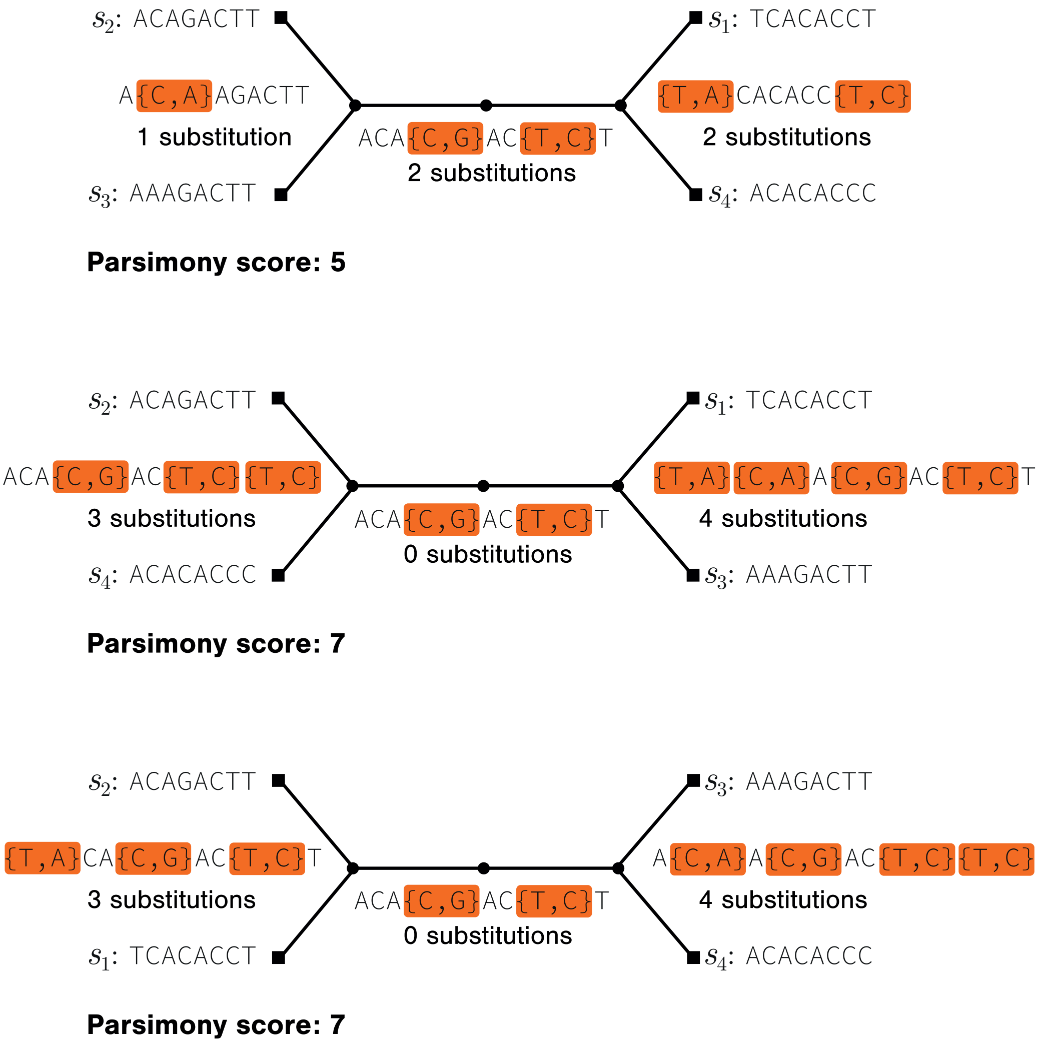


Figure 6.14: Calculation of parsimony scores for the three unrooted trees on four tips. The dot on the internal branch is the artificially added root (the first step in the Fitch algorithm).

6.3.2.6 Statistical inconsistency of the parsimony method

A principle shortcoming of the parsimony method is that — by definition — it always assumes the smallest number of substitutions possible, even in cases where more substitutions might

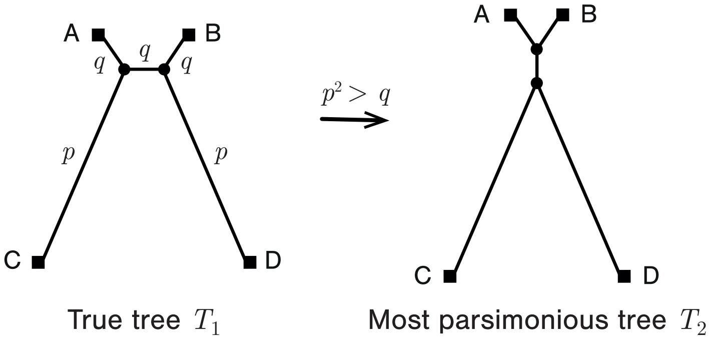


Figure 6.15: Long branch attraction. When the correct tree (T_1) has two long branches separated by a short internal branch, parsimony tends to reconstruct a wrong tree (T_2) with the two long branches grouped together. This phenomenon is called long branch attraction.

provide a more likely explanation of the data. One specific side effect of this is that parsimony does not acknowledge the existence of back-substitutions. For example, consider a site that has an A at both ends of a branch. Although there may have been several hidden substitutions along the branch, such as A → G → A, parsimony always assumes no substitution at all in such a case. The result of this assumption is a particular form of bias called *long branch attraction*. This is known to arise when two very long branches are connected through a relatively short internal branch as seen in Figure 6.15, left.

To more carefully illustrate the problem, assume that we have the states 0 and 1 for each site in the MSA (rather than A, G, C, and T), and let the probability of observing a change on the long branches in the true tree T_1 (Figure 6.15) (leading to C and D) be p, and on the three short branches be q.

Now assume sequences evolved on the tree T_1. Approximately, if $p^2 > q$, meaning convergent changes on the two long branches are more likely than a single change on the short internal branch, then C and D are more often in the same state (and different from A and B) than A and C are in the same state (and different from B and D). In turn, parsimony reconstruction based on the sequences that evolved on tree T_1 will put C and D into a cherry to minimise the number of substitutions (tree T_2). When the number of sites goes to infinity, parsimony will necessarily pick the wrong tree T_2.

While roughly correct, this argument ignores the probabilities of no change on a branch and scenarios with changes on the branches leading to A and B. Properly considering all possibilities, the condition for parsimony selecting the wrong tree in our example when the number of sites goes to infinity is $p^2 > q - q^2$. This is shown rigorously in Felsenstein (1978) and Felsenstein (2003). Thus, the parsimony methods are statistically inconsistent even under such simple evolutionary models.

This statistical inconsistency means that parsimony is rarely used as a principal phylogenetic inference method in macroevolution nowadays, though it is still popular within parts of the cladist community. However, the parsimony concept continues to be useful in fields such as epidemiology or developmental biology. For example, the parsimony concept has been successfully employed to rapidly place additional tips on huge pathogen trees to avoid rerunning analyses when new data become available Turakhia et al. (2021). Further, particular settings exist where a sequence site is known to usually only change once, such as in certain experiments in developmental biology. In such settings, the parsimony concept is applicable again (see Chapter 12 for more details).

6.3.3 Probabilistic approach: maximum likelihood methods

The *probabilistic approaches* (dating to Edwards and Cavalli-Sforza (1964)) assume an explicit probabilistic model of evolution underlying the data. It evaluates the likelihood of the model parameters θ given the data D, $L(\theta; D) = P(D|\theta)$ (see also Box 25 on page 116 for the likelihood function), and then compares the likelihood for different parameters. In the *maximum likelihood (ML)* approach, the parameters maximising the likelihood function are estimated and reported as maximum likelihood parameter estimates. In the *Bayesian approach*, the posterior distribution of parameters, which is a function of the likelihood, is estimated (see Chapter 10 for details). Due to short-comings of the distance-based methods (only pairwise distances are considered; higher-order relationships are ignored) and the parsimony method (long branch attraction), maximum likelihood methods, together with Bayesian phylogenetic methods (Chapter 10), are the methods of choice for most phylogenetic studies.

In a phylogenetic context, the probabilistic model consists of two components. The first component of the model is the tree with branch lengths, $\mathcal{T}$, which describes how the studied biological unit (such as species, viruses, single cells, and so on) replicated. The tree with branch lengths may be a parameter or may be generated under some probabilistic tree-generating model (see Chapter 9 for tree-generating models). In this section, we assume that the tree is a parameter of the model. The second component is a model of sequence evolution that describes how nucleotides (or codons or amino acids) change over time with a rate matrix Q. Commonly used models are, for example, JC69, HKY and GTR (see Chapter 5 for details on substitution models). The probabilistic model gives rise to the probability of the data, the MSA D, given the tree and rate matrix, $P(D|\mathcal{T}, Q)$. The likelihood of $\mathcal{T}$ and Q is thus $L(\mathcal{T}, Q; D) = P(D|\mathcal{T}, Q)$. For a given probabilistic model, we aim to find the tree and rate matrix that maximise the likelihood $\max_{\mathcal{T},Q} L(\mathcal{T}, Q; D)$. This optimal tree and rate matrix are the maximum likelihood estimates.

We will now explain how to calculate $P(D|\mathcal{T}, Q)$, the *phylogenetic likelihood*.

In a very naive approach, we could calculate $P(D|\mathcal{T}, Q)$ by simulation. A probabilistic description of the process means that given the parameters $\mathcal{T}$ and Q, we can simulate sequence data along the tree and obtain a simulated MSA of n sequences. The probability $P(D|\mathcal{T}, Q)$ is the frequency with which the particular MSA D will be simulated along the given tree for

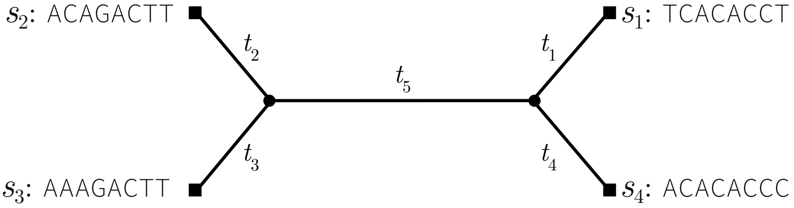


Figure 6.16: Unrooted tree with the sequences at the tips used as an example for the likelihood calculation. Again, this is the tree obtained using UPGMA (without the root).

the given rate matrix when simulating many MSAs. This simulation-based approach is a very slow way to determine $P(D|\mathcal{T}, Q)$, as we have to simulate many MSAs.

We next discuss how to analytically calculate the probability of the MSA given $\mathcal{T}$ and Q. We note that the probability of the MSA (the sequences at the tips) is the joint probability of the sequences at the tips and sequences at all internal nodes of the tree, summed over all possible sequences at internal nodes.

In what follows, we will consider nucleotide MSAs, but the approaches for codon or amino acid sequence MSAs are equivalent. To further illustrate how the probability is calculated based on a given tree, we will again use the toy MSA and the UPGMA tree as displayed in Figure 6.16.

6.3.3.1 Calculating the likelihood using a brute-force approach

We will now show how to calculate the probability of the sequences at the tips for a given tree and rate matrix by explicitly summing over all internal node sequences. Typically, and in what follows, we assume one of the time-reversible models introduced in Chapter 5. These models all assume that the sites evolve independently from one another. Consequently, we can calculate the probability for each site separately and then take the product over the single-site probabilities to get the full probability. Put into mathematical equations, let us assume that the MSA consists of n sequences $(s_1, \dots, s_n)$ with m sites each, then the probability of the sequences is

$$P(s_1, \dots, s_n|\mathcal{T}, Q) = \prod_{j=1}^{m} P(s_{1,j}, \dots, s_{n,j}|\mathcal{T}, Q), \tag{6.8}$$

where $s_{k,j}$ is the jth site of sequence s_k.

To evaluate $P(s_{1,j}, \dots, s_{n,j}|\mathcal{T}, Q)$ for all j, we add a root to the unrooted tree at an arbitrary position, leading to tree $\mathcal{T}_r$ (node s_7 in Figure 6.17). In what follows, we calculate the probability for the rooted tree $P(s_{1,j}, \dots, s_{n,j}|\mathcal{T}_r, Q)$ (we will later discuss the impact of this

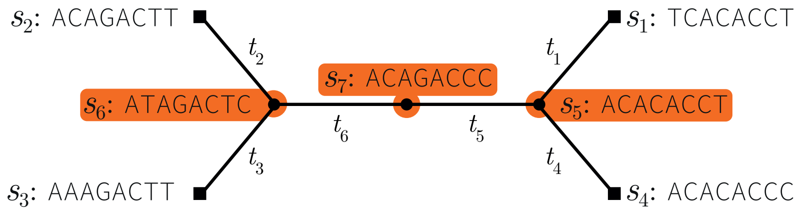


Figure 6.17: Unrooted tree with sequences at the tips for which we calculate the probability of the sequences at the tips. The root node position and the sequences at the internal nodes have been assigned arbitrarily.

root placement on the probability). Next, we assign arbitrary sequences to all internal nodes (nodes s_5, s_6 and s_7 in Figure 6.17).

For n sequences, the rooted tree has $n-1$ internal nodes, with sequences $s_{n+1}, \ldots, s_{2n-1}$. By summing over the possible nucleotides at the internal nodes, we obtain:

$$P(s_{1,j}, \ldots, s_{n,j} | \mathcal{T}_r, Q) = \sum_{s_{n+1,j} \in \{\mathsf{T,C,A,G}\}} \cdots \sum_{s_{2n-1,j} \in \{\mathsf{T,C,A,G}\}} P(s_{1,j}, s_{2,j}, \ldots, s_{2n-1,j} | \mathcal{T}_r, Q). \tag{6.9}$$

The probability $P(s_{1,j}, s_{2,j}, \ldots, s_{2n-1,j} | \mathcal{T}_r, Q)$ is computed by

1. calculating the transition probability $p_{s_{l_1,j}, s_{l_2,j}}(t_l)$ from the ancestral node to the descendant node for each branch l with starting sequence s_{l_1}, ending sequence s_{l_2}, and branch length t_l, and
2. calculating the probability of the nucleotide at the root using the equilibrium probability, $\pi(s_{2n-1,j})$.

We get the overall expression:

$$P(s_{1,j}, s_{2,j}, \ldots, s_{2n-1,j} | \mathcal{T}_r, Q) = \pi(s_{2n-1,j}) \prod_{l=1}^{2n-2} p_{s_{l_1,j}, s_{l_2,j}}(t_l). \tag{6.10}$$

For example, for site $j=2$ in the example of Figure 6.17, the equation is:

$$P(s_{1,2}, s_{2,2}, \ldots, s_{2n-1,2} | \mathcal{T}_r, Q) = \pi_{\mathsf{C}} p_{\mathsf{C,C}}(t_5) p_{\mathsf{C,C}}(t_4) p_{\mathsf{C,C}}(t_1) p_{\mathsf{C,T}}(t_6) p_{\mathsf{T,C}}(t_2) p_{\mathsf{T,A}}(t_3). \tag{6.11}$$

By summing over all internal sequences and multiplying across all sites, we obtain the probability of the MSA given a rooted tree. As we assume time-reversible models, the model does not distinguish the direction of time flow along the branches. This means that we obtain the same probability regardless of where the root is added to the unrooted tree (for a formal explanation, see Section 6.4.1). In summary,

$$P(s_1,\ldots,s_n|\mathcal{T},Q)=\prod_{j=1}^{m}\sum_{s_{n+1,j}\in\{\mathtt{T},\mathtt{C},\mathtt{A},\mathtt{G}\}}\cdots\sum_{s_{2n-1,j}\in\{\mathtt{T},\mathtt{C},\mathtt{A},\mathtt{G}\}}\pi(s_{2n-1,j})\prod_{l=1}^{2n-2}p_{s_{l_1,j},s_{l_2,j}}(t_l). \tag{6.12}$$

We now assess the runtime of this approach for calculating the MSA probability and thus the likelihood of $\mathcal{T}$ and Q given the MSA. In our example, we have three internal nodes, which can have one of the four nucleotides at site j. Thus, we have $4\times 4\times 4 = 64$ possibilities for the nucleotides at site j, and the sum in our example consists of 64 terms. In general, the sum over all nucleotides at site j in a tree with n tips has 4^{n-1} terms, as the tree has $n-1$ internal nodes. Furthermore, for each nucleotide configuration, the approach considers each branch in the tree ($2n-2$ branches). We finally need to multiply the likelihood over all m sites in the MSA. Overall, the runtime of this brute-force algorithm is $\mathcal{O}(m4^n n)$, meaning it is exponential in n.

6.3.3.2 Calculating the likelihood using Felsenstein's pruning algorithm

A more efficient way to calculate the likelihood is called *Felsenstein's pruning algorithm* (Felsenstein 1973; Felsenstein 1981). The likelihood of the tree and rate matrix given an MSA can be computed in linear time given that the transition probabilities $p_{X,Y}(t)$ are known. Analogously to the brute-force algorithm, we arbitrarily root the tree and independently calculate the likelihood for each site in the MSA. However, we now sum over the possible nucleotides at the internal nodes in the tree more efficiently using dynamic programming (the concept of dynamic programming was introduced in Chapter 3). The strategy of recursively traversing the tree in Felsenstein's pruning algorithm is analogous to the strategy in the Fitch algorithm (Section 6.3.2.3).

We now show how to calculate the sequence probability given a rooted tree and rate matrix for a single site in the MSA. Given a nucleotide $X\in\{\mathtt{T},\mathtt{C},\mathtt{A},\mathtt{G}\}$ at node k, let the probability of the nucleotides at the tips descending from node k be $P(D_k|X)$. This probability is central to the Felsenstein's pruning algorithm. For a tip node k with nucleotide Y, we have $P(D_k|X)=1$ if $X=Y$ and $P(D_k|X)=0$ otherwise. If the site at tip k is a gap (a - in the MSA), we initialise with $P(D_k|X)=1$ for $X\in\{\mathtt{T},\mathtt{C},\mathtt{A},\mathtt{G}\}$, meaning that we assume any nucleotide was possible at that site (and in particular we assume it is not a real gap). Next, let k be an internal node with the descending adjacent nodes l and m and branch lengths t_l, t_m, for which we already calculated $P(D_l|Y)$, $P(D_m|Z)$ for $Y,Z\in\{\mathtt{T},\mathtt{C},\mathtt{A},\mathtt{G}\}$. The probability $P(D_k|X)$ is obtained by multiplying the probabilities of the two descendant subtrees and the transition probabilities from k to l and m,

$$P(D_k|X) = \left(\sum_{Y \in \{\mathsf{T},\mathsf{C},\mathsf{A},\mathsf{G}\}} p_{X,Y}(t_l) P(D_l|Y) \right) \times \left(\sum_{Z \in \{\mathsf{T},\mathsf{C},\mathsf{A},\mathsf{G}\}} p_{X,Z}(t_m) P(D_m|Z) \right). \quad (6.13)$$

In summary, after defining the probabilities at the tips, we prune cherries recursively towards the root using the formula for $P(D_k|X)$ defined in Equation (6.13). The pruning terminates at the root r, where we calculate $P(D_r|X)$, where $X \in \{\mathsf{T}, \mathsf{C}, \mathsf{A}, \mathsf{G}\}$. Finally, the probability of the sequences observed at the tips at site j is obtained by summing over the four possible root nucleotides:

$$P(s_{1,j}, \ldots, s_{n,j}|\mathcal{T}, Q) = \sum_{X \in \{\mathsf{T},\mathsf{C},\mathsf{A},\mathsf{G}\}} P(D_r|X)\pi_X. \quad (6.14)$$

For an MSA of length m on the tree shown in Figure 6.18, let us compare the brute-force way of writing the likelihood,

$$\begin{aligned} P(s_1, s_2, s_3, s_4|\mathcal{T}, Q) = \prod_{j=1}^{m} \sum_{s_{7,j} \in \{\mathsf{T},\mathsf{C},\mathsf{A},\mathsf{G}\}} \sum_{s_{6,j} \in \{\mathsf{T},\mathsf{C},\mathsf{A},\mathsf{G}\}} \sum_{s_{5,j} \in \{\mathsf{T},\mathsf{C},\mathsf{A},\mathsf{G}\}} & \pi(s_{7,j}) p_{s_{7,j},s_{6,j}}(t_6) \\ \times\, p_{s_{6,j},s_{3,j}}(t_3) p_{s_{6,j},s_{2,j}}(t_2) p_{s_{7,j},s_{5,j}}(t_5) p_{s_{5,j},s_{4,j}}(t_4) & p_{s_{5,j},s_{1,j}}(t_1), \end{aligned} \quad (6.15)$$

with Felsenstein's likelihood,

$$\begin{aligned} P(s_1, s_2, s_3, s_4|\mathcal{T}, Q) = \prod_{j=1}^{m} \sum_{s_{7,j} \in \{\mathsf{T},\mathsf{C},\mathsf{A},\mathsf{G}\}} & \pi(s_{7,j}) \\ & \times \left(\sum_{s_{6,j} \in \{\mathsf{T},\mathsf{C},\mathsf{A},\mathsf{G}\}} p_{s_{7,j},s_{6,j}}(t_6) p_{s_{6,j},s_{3,j}}(t_3) p_{s_{6,j},s_{2,j}}(t_2) \right) \\ & \times \left(\sum_{s_{5,j} \in \{\mathsf{T},\mathsf{C},\mathsf{A},\mathsf{G}\}} p_{s_{7,j},s_{5,j}}(t_5) p_{s_{5,j},s_{4,j}}(t_4) p_{s_{5,j},s_{1,j}}(t_1) \right). \end{aligned} \quad (6.16)$$

Notice how the summation signs moved "down the tree" when we used Felsenstein's algorithm.

We now illustrate the pruning algorithm with an example. For the example tree in Figure 6.18, where we consider a single nucleotide position, we state the values of $P(D_k|X)$ for all tip nodes within the figure. We now need to calculate the probabilities for each nucleotide marked by ?.

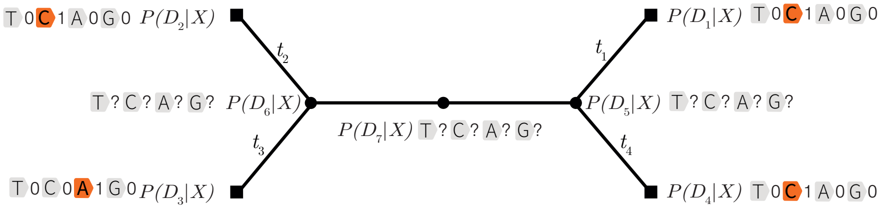

T 0 C 1 A 0 G 0 $P(D_2|X)$
t_2
T ? C ? A ? G ? $P(D_6|X)$
t_3
T 0 C 0 A 1 G 0 $P(D_3|X)$
$P(D_7|X)$ T ? C ? A ? G ?
$P(D_1|X)$ T 0 C 1 A 0 G 0
t_1
$P(D_5|X)$ T ? C ? A ? G ?
t_4
$P(D_4|X)$ T 0 C 1 A 0 G 0

Figure 6.18: Example for the setup of the likelihood computation using Felsenstein's pruning algorithm. For a node k, the number to the right of a nucleotide states the probability $P(D_k|X)$ with X corresponding to that nucleotide. The orange nucleotides at the tips represent the data.

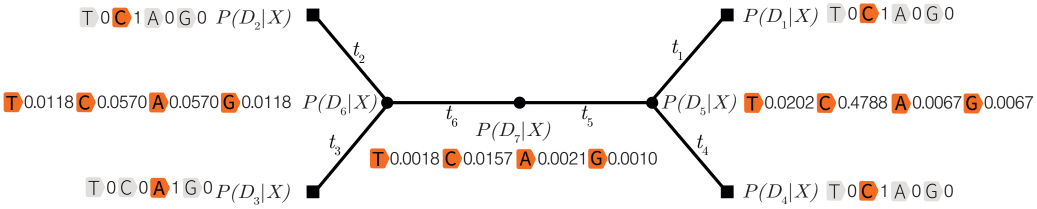

T 0 C 1 A 0 G 0 $P(D_2|X)$
t_2
T 0.0118 C 0.0570 A 0.0570 G 0.0118 $P(D_6|X)$
t_3
T 0 C 0 A 1 G 0 $P(D_3|X)$
t_6
$P(D_7|X)$
T 0.0018 C 0.0157 A 0.0021 G 0.0010
t_5
$P(D_1|X)$ T 0 C 1 A 0 G 0
t_1
$P(D_5|X)$ T 0.0202 C 0.4788 A 0.0067 G 0.0067
t_4
$P(D_4|X)$ T 0 C 1 A 0 G 0

Figure 6.19: Example of the likelihood computation using Felsenstein's pruning algorithm. The orange nucleotides at the tips represent the data. The question marks ? in Figure 6.18 were evaluated assuming a K80 nucleotide substitution model with $\kappa = 2$ (Section 5.3.2), and branch lengths $t_1 = t_2 = \ldots = t_6 = 0.1$.

We can calculate $P(D_6|\mathsf{T})$ as follows:

$$\begin{aligned} P(D_6|\mathsf{T}) &= \sum_{Y \in \{\mathsf{T},\mathsf{C},\mathsf{A},\mathsf{G}\}} p_{\mathsf{T},Y}(t_2) P(D_2|Y) \times \sum_{Z \in \{\mathsf{T},\mathsf{C},\mathsf{A},\mathsf{G}\}} p_{\mathsf{T},Z}(t_3) P(D_3|Z) \\ &= p_{\mathsf{T},\mathsf{A}}(t_2) \times p_{\mathsf{T},\mathsf{C}}(t_3). \end{aligned} \tag{6.17}$$

A complete example of Felsenstein's pruning algorithm with the internal node probabilities calculated for each nucleotide is shown in Figure 6.19.

6.3.3.3 Time complexity and statistical consistency of maximum likelihood

Next, we determine the time complexity of Felsenstein's pruning algorithm. In each pruning step, we sum over four states twice (the two descending branches and four possible nucleotides at the ends of these branches), meaning we perform a constant number of operations. The number of pruning steps equals the number of internal nodes in the rooted tree, and thus, the full pruning procedure has the runtime of $\mathcal{O}(n)$. In addition, this procedure has to be performed for each of the m sites. Thus, in total, the runtime of the algorithm is $\mathcal{O}(nm)$ which is linear in n (as opposed to $\mathcal{O}(m4^n n)$ for the brute-force approach).

While we can obtain the likelihood of a tree for a given MSA in linear time using Felsenstein's pruning algorithm, inferring the maximum likelihood phylogeny is NP-hard (Roch 2006) as we essentially have to consider every unrooted tree on n tips for an MSA of n sequences.

Maximum likelihood tree reconstruction is statistically consistent (see e.g. Felsenstein (1973) and Felsenstein (2003)), which means the true tree is returned when the method is presented with an infinite amount of data in the form of infinitely long sequences (recall Box 27 on page 152 on statistical consistency).

Maximum likelihood methods have been criticised by cladists, arguing they make too many assumptions on the details of the evolutionary process; see, for example, Farris (1983), and Felsenstein (2003) for more details and more references. However, it has been shown that the parsimony tree is, in fact, equal to the maximum likelihood tree when assuming a no-common-mechanism substitution model (Tuffley and Steel 1997). This model makes many assumptions and is highly over-parameterised, as each site may have its own rate for each branch in the tree, meaning it is much more detailed than the common substitution models introduced in Chapter 5.

6.4 From unrooted trees to time trees

6.4.1 Time-reversibility implies that differently rooted trees have the same likelihood

In Section 6.3.2, we outlined that the parsimony approach cannot distinguish between different rooted trees based on the same unrooted tree. Likelihood methods assuming time-reversible substitution models cannot distinguish between different rooted trees that are based on the same unrooted tree: In calculating the likelihood, the assumption of time-reversibility implies that the likelihood remains the same no matter where one places the root of the tree. We illustrate this in a simple example of a two-tip tree (Figure 6.20) where we consider two possible roots, D_1 and D_2.

We consider a single site that has nucleotides s_1 and s_2 at the two tips of the tree; $s_i \in \mathcal{N} = \{\mathtt{T}, \mathtt{C}, \mathtt{A}, \mathtt{G}\}$. The likelihood of the tree with root D_1 (Figure 6.20, left) is then

$$P(D_1) = \sum_{X \in \mathcal{N}} \pi_X p_{X,s_1}(t_1) p_{X,s_2}(t_2 + t_3). \tag{6.18}$$

The transition probability of going from the internal state X to the observed nucleotide s_2 in time $t_2 + t_3$ can be divided into the transition probability to another internal state (at node

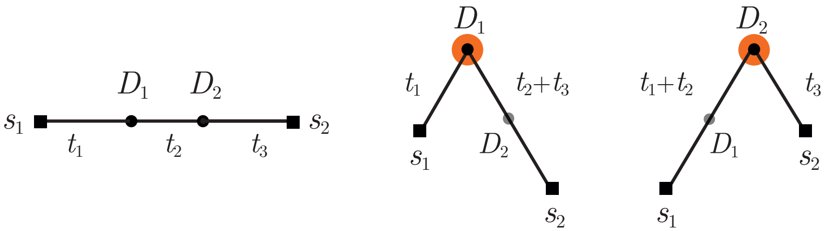


Figure 6.20: Illustration of the role of time reversibility in the context of likelihood calculations.

D_2; Figure 6.20, left) and then to s_2. As we do not know which state the nucleotide is in at node D_2, we have to sum over all possibilities:

$$p_{X,s_2}(t_2+t_3) = \sum_{Y\in\mathcal{N}} p_{X,Y}(t_2)p_{Y,s_2}(t_3). \tag{6.19}$$

We can now substitute this into Equation (6.18) and rewrite these probabilities as

$$\begin{aligned}
P(D_1) &= \sum_{X\in\mathcal{N}} \pi_X p_{X,s_1}(t_1) \sum_{Y\in\mathcal{N}} p_{X,Y}(t_2)p_{Y,s_2}(t_3) \\
&= \sum_{X\in\mathcal{N}} \sum_{Y\in\mathcal{N}} \pi_X p_{X,s_1}(t_1)p_{X,Y}(t_2)p_{Y,s_2}(t_3) \\
&= \sum_{X\in\mathcal{N}} \sum_{Y\in\mathcal{N}} \underbrace{\pi_X p_{X,Y}(t_2)}_{\overset{(5.42)}{=}\pi_Y p_{Y,X}(t_2)} p_{X,s_1}(t_1)p_{Y,s_2}(t_3) \\
&= \sum_{Y\in\mathcal{N}} \pi_Y p_{Y,s_2}(t_3) \sum_{X\in\mathcal{N}} p_{Y,X}(t_2)p_{X,s_1}(t_1) \\
&= \sum_{Y\in\mathcal{N}} \pi_Y p_{Y,s_2}(t_3)p_{Y,s_1}(t_1+t_2) \\
&= P(D_2).
\end{aligned} \tag{6.20}$$

Thus, we prove that the likelihoods for both rooted trees are the same regardless of where the root is placed in the tree.

Applied to trees with more than two tips, this reasoning shows that the likelihood is independent of the position of the root along an edge. Assuming continuity of the likelihood as the root moves past internal nodes, we further obtain that the likelihood must also remain fixed regardless of where we place the root on the tree. This property of phylogenetic likelihoods under reversible substitution models has been referred to as the *pulley principle* and is discussed further by Felsenstein (1981).

6.4.2 Rooting the tree

As discussed, core phylogenetic inference algorithms presented in this chapter return an unrooted tree. We can *root the tree* by including an *outgroup* into the analysis. An outgroup is a group of individuals/species distant from the rest of the data (the ingroup) considered in the tree. For example, for mammals, one could use bird sequences as an outgroup; for the transmission tree of HIV subtype B, one could use sequences from HIV subtype C or A. The point where the outgroup connects to the phylogeny of the species of interest is defined as the root of the ingroup. Thus, one assumes *a priori* that the outgroup attaches to the root of the remaining phylogeny because it is assumed to have diverged from the phylogeny of interest at a much earlier point in time.

6.4.3 Adding a calendar time scale to rooted phylogenetic trees

The tree inference methods discussed so far estimate trees with branch lengths representing the number of substitutions along a branch. It is useful to have trees with branch lengths and internal node times corresponding to calendar time for many applications: for instance, to time speciation and extinction events on the tree of life or to estimate the rate of spread of an epidemic. Such time-scaled trees are called *time trees*.

To obtain branch lengths in calendar time from branch lengths in the number of substitutions, we require a *clock rate* parameter that quantifies the expected number of substitutions per calendar time unit. This clock rate is typically unknown and has to be co-estimated with the calendar time branch lengths. The clock rate may be the same across all branches (strict molecular clock (Zuckerkandl and Pauling 1962)) or may vary across branches (relaxed molecular clock (Drummond et al. 2006)). For simplicity, we now consider a strict clock, but the general arguments also hold for a relaxed clock model.

If all tips are sampled at the same time and no additional information is provided aside from the alignments, we can only infer branch lengths proportional to calendar time but not the absolute branch lengths. For illustration, suppose we have a tree of height t and a clock rate r. Scaling all branches in the tree by 2 (obtaining a new tree with height $2t$) and shrinking the clock rate to $^{r}/_{2}$ explains the data equally well; we simply slowed down the whole process (see also Figure 6.21 and Section 5.3.5).

This correlation can also be seen in Felsenstein's pruning likelihood calculation, as the rate and time always appear as a product (labelled d in Section 5.3.5). Thus, absolute time and the clock rate are not identifiable together; thus, we cannot estimate absolute time unless we include further information.

When sequences evolve according to a strict molecular clock and are sampled at the same time, the inferred UPGMA tree has branch lengths that are directly proportional to calendar time (see Theorem 6.3.1), meaning we can estimate relative calendar time (but not absolute time).

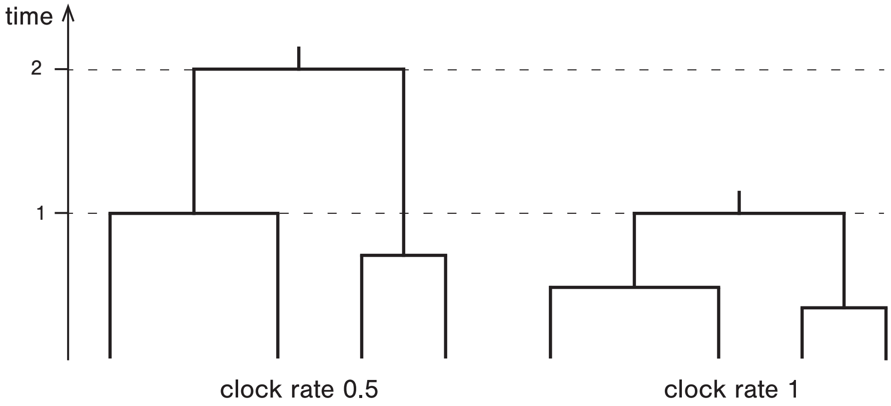


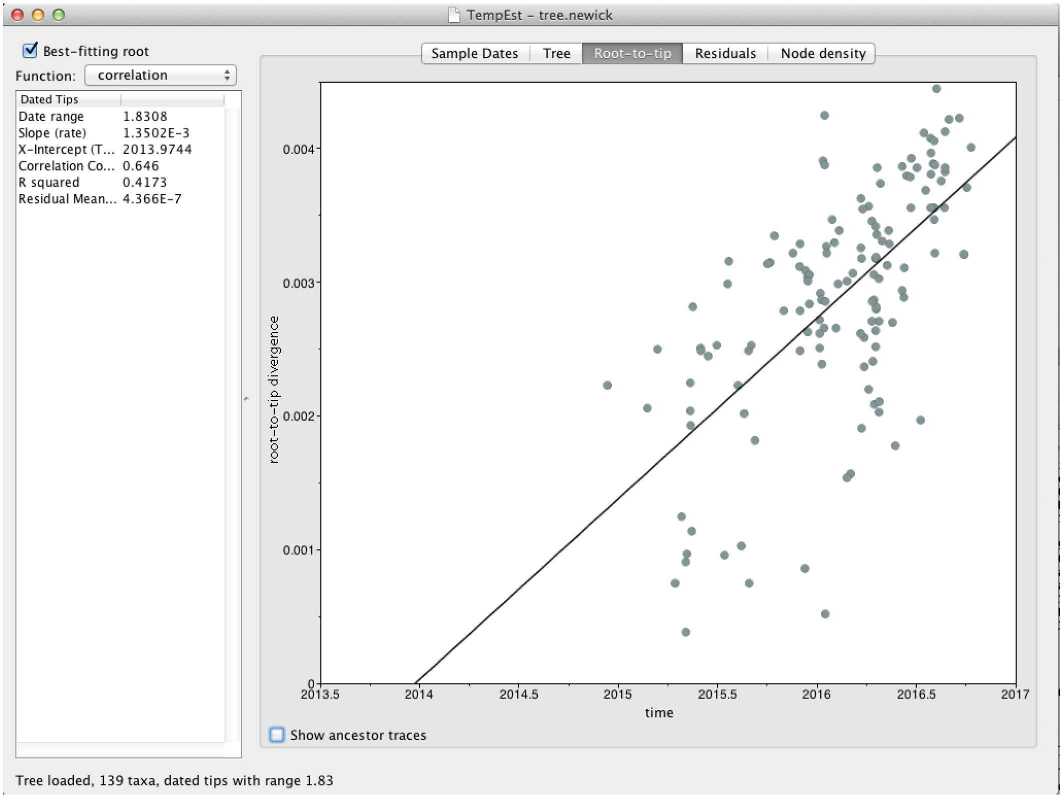
Figure 6.21: Illustration of two combinations of dated topology and clock rate values that have the same phylogenetic likelihood given the present-day sequence data. It is impossible to differentiate between them from present-day sequence data alone. Data from the past, such as fossils or ancient sequences, can resolve this non-identifiability.

Scaling branch lengths for the other presented methods involves rooting the tree (e.g. through an outgroup) first. Fossils provide information on when the most recent common ancestor of certain species existed for obtaining time estimates. They can be used in the inference to inform the calendar time of specific nodes, which constrains the range of possible clock rates. If enough fossils are included, a precise estimate for the clock rate can be inferred (Heath 2012).

If we sample sequences from multiple time points, we can estimate the clock rate from the sequences and sampling times, given that sufficient evolution occurred along the branches. To provide an intuition, for a rooted tree with branch lengths in the number of substitutions and under the assumption of a strict clock, we can plot the calendar time of a sample on the x-axis and the sum of branch lengths from the root to the particular sample on the y-axis (Figure 6.22). Later samples will have a larger y-value, and the regression slope through the data points is an estimator for the clock rate (Rambaut et al. 2016).

A computationally efficient way to extract this clock signal from serially sampled phylogenies with branch lengths in the number of substitutions and to estimate the clock rate and date the phylogeny in calendar time is the *least-squares dating (LSD) method* (To et al. 2016). This method uses a binary phylogenetic tree with known branch lengths in the number of substitutions — inferred using a distance, parsimony, or maximum likelihood-based approach — and information about the sampling times of the tips to estimate the clock rate and the dates of all internal nodes.

In what follows, we explain the LSD method. Suppose the input consists of a rooted binary tree R on n sequences, where internal nodes of R are numbered $1, \ldots, n-1$ and leaves



TempEst – tree.newick

☑ Best-fitting root

Function: correlation

Dated Tips	
Date range	1.8308
Slope (rate)	1.3502E-3
X-Intercept (T...	2013.9744
Correlation Co...	0.646
R squared	0.4173
Residual Mean...	4.366E-7

☐ Show ancestor traces

Tree loaded, 139 taxa, dated tips with range 1.83

Figure 6.22: Correlation between the sampling times of different Zika sequences and their distance to the root in terms of substitutions, shown in the software TempEst (`http://tree.bio.ed.ac.uk/software/tempest/`) (Rambaut et al. 2016) using data from Bošková, Stadler and Magnus (2018). The slope of the regression provides an estimate of the clock rate.

$n, \ldots, 2n-1$. The direct ancestor of node i is denoted by $a(i)$. Sampling times in calendar time are denoted by $t_n, \ldots, t_{2n-1}$. The internal node calendar times (to be estimated) are denoted by $t_1, \ldots, t_{n-1}$. The clock rate (to be estimated) is denoted by r. The model then assumes that the branch lengths in the number of substitutions (b_i) are the result of a strict molecular clock with a constant clock rate acting over a given calendar time interval $\Delta t = t_i - t_{a(i)}$, together with a Gaussian noise term $\epsilon_i \sim \text{Normal}(0, \sigma_i^2)$ stemming from sampling and estimation errors:

$$b_i = r\left(t_i - t_{a(i)}\right) + \epsilon_i. \tag{6.21}$$

The estimates of the global clock rate and the internal node dates are obtained by minimising the weighted least squares criterion:

$$\Phi(r, t_1, \ldots, t_{n-1}) = \sum_{i=2}^{2n-1} \frac{1}{\sigma_i^2} \left(b_i - r(t_i - t_{a(i)})\right)^2 . \tag{6.22}$$

Here, the variance terms σ^2 are unknown, but using the Poisson nature of the substitution process, one can arrive at the following assumptions:

$$\sigma_i^2 = \frac{r(t_i - t_{a(i)})}{m}, \tag{6.23}$$

and

$$\hat{\sigma_i^2} = \frac{b_i + {}^{c}/_{m}}{m}, \tag{6.24}$$

where c is a constant, added to avoid infinite weights in the case of zero branch lengths ($b_i = 0$), and m is the sequence length.

This method has been shown to be reasonably accurate even in cases with minor uncorrelated violations of the strict molecular clock, as these are absorbed into the added Gaussian noise term (To et al. 2016). With rooted input trees, the LSD algorithm can be implemented with approximately linear time complexity ($\mathcal{O}(n)$, where n is the number of tips). When the method is extended to account for unrooted trees, the time complexity is approximately quadratic ($\mathcal{O}(n^2)$).

6.5 Searching the tree space

Tree reconstruction based on distance-based optimality methods, parsimony, or maximum likelihood methods is NP-hard and requires a search among all possible unrooted trees. Additionally, all possible branch lengths must be checked for optimality for all methods except parsimony. For each tree with branch lengths, the sum of squares, the parsimony score, or the likelihood value is evaluated, and the tree with the best score is returned.

The search through branch lengths requires optimising a continuous function (least squares function or likelihood function) over $2n - 3$ real variables, namely the $2n - 3$ branch lengths for a proposed unrooted tree on n tips. Such optimisation can be done with the commonly used *hill climbing* concept (see e.g. Yang (2014, Chapter 4.6)). The hill climbing algorithm starts with some initial branch lengths. Then, new branch lengths are proposed and accepted if they increase the likelihood function. The algorithm terminates if no increase can be found.

Searching the space of all tree topologies (trees without branch lengths) is more complicated, as the space of trees is a huge discrete space. Checking each tree for optimality in tree space will be too slow (see again Table 6.1); thus, we generally resort to heuristics that attempt to check as many trees as possible. Note that this means that we may miss the best tree. However,

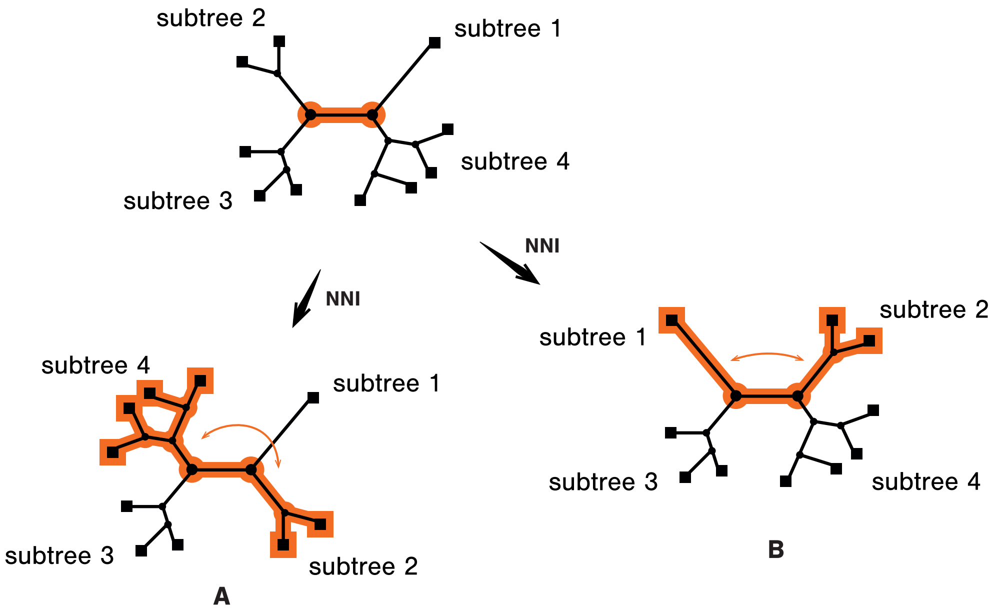


Figure 6.23: The nearest-neighbour interchange (NNI) move. Each internal branch in the tree connects four subtrees or nearest neighbours. Exchanging a subtree on one side of the branch with another on the other side constitutes an NNI move. Two such rearrangements are possible for each internal branch (here, **A** subtree 2 $\leftrightarrow$ subtree 4, and **B** subtree 2 $\leftrightarrow$ subtree 1).

finding the optimal tree is computationally intractable unless $P = NP$ or computational speed increases drastically.

Checking random trees from all possible trees is inefficient, as we often check trees with very bad scores. Thus, chances are high that we will miss the best tree unless we check many trees. A better approach uses a random walk to search the tree space. The method starts with an arbitrary tree and iteratively modifies the current tree by replacing it with a similar tree with a better score. This is the same idea as hill climbing, but in this case, applied to the discrete space of tree topologies. This procedure requires specific *tree moves* that propose new trees, three of which — NNI, SPR, TBR — we will briefly discuss here (for more details, see, for example, Felsenstein (2003) and Yang (2014)).

The *NNI* move, or ***N**earest-**N**eighbour **I**nterchange*, switches two neighbouring subtrees. In particular, it chooses an internal branch uniformly at random. This branch has four subtrees attached. Two subtrees separated by the internal branch are then exchanged, as shown in Figure 6.23.

The other two moves are *SPR*, ***S**ubtree **P**runing and **R**egrafting*, and *TBR*, ***T**ree **B**isection and*

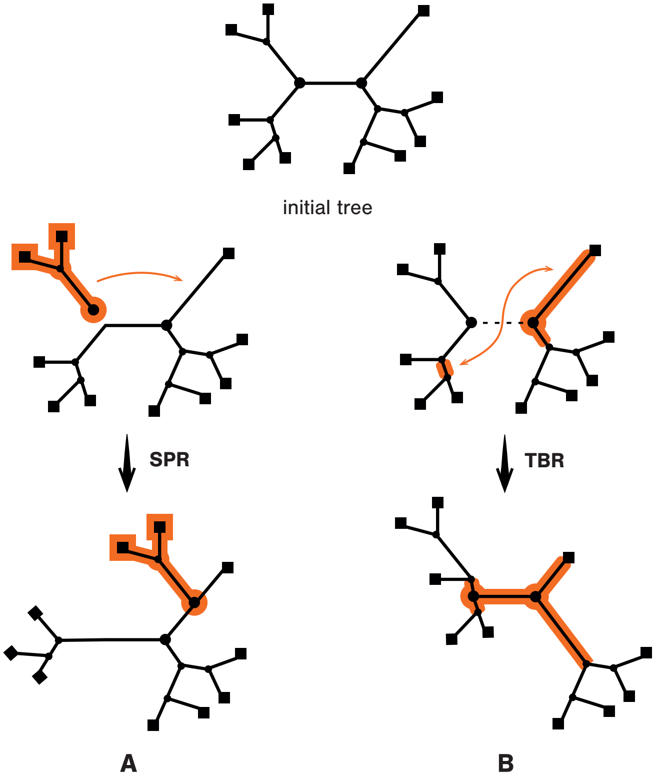


Figure 6.24: **A** Tree move by subtree pruning and regrafting (SPR). A subtree is detached and then reattached to a different location on the tree. **B** Tree move by tree bisection and reconnection (TBR). The tree is broken into two subtrees by removing an internal branch. Two branches, one from each subtree, are then chosen uniformly at random and rejoined to form a new tree.

Reconnection. In SPR, an internal branch with its descending subtree is first chosen uniformly at random. The chosen branch and subtree are detached from the remaining tree. Then, a branch in the remaining tree is chosen uniformly at random to which the detached branch and subtree are regrafted (Figure 6.24 A). In TBR, a random internal branch is chosen and deleted, splitting a tree into two unrooted subtrees. Two branches, one in each subtree, are chosen uniformly at random and merged to reconnect the two subtrees, as shown in Figure 6.24 B.

State-of-the-art implementations employing efficient heuristics (including the moves described above) allow us to perform maximum likelihood tree inference for datasets containing thousands of sequences. Some of the commonly used software packages are PhyML (http://www.atgc-montpellier.fr/phyml/) (Guindon and Gascuel 2003) and RaxML (http://embnet.vital-it.ch/raxml-bb/) (Stamatakis 2006). RaxML is optimised for large datasets and can infer trees on the order of 10^5 tips.

In Chapter 7, we will discuss how to pick an appropriate substitution model for maximum likelihood inference and how to acknowledge uncertainty in estimation results; the overall maximum likelihood approach is summarised in Box 29 on page 198.

6.6 Examples of applications of phylogenetic reconstruction methods

6.6.1 The first phylogenies

Michener and Sokal (1957) reconstructed the first phenetic tree (Figure 8 in their paper) based on bee data in 1957. The tree was built based on morphological traits only.

Edwards and Cavalli-Sforza (1964) reconstructed the first cladistic (parsimony) tree in 1964. The researchers explored the evolution of human populations using the blood groups and their frequencies. Interestingly, the tree of blood type frequencies is in line with what is known about the migration of human populations. The same paper also introduced the maximum likelihood method, but since the pruning approach was not known, the method was computationally very slow.

Maximum likelihood methods were widely employed upon publication of the pruning algorithm (Felsenstein 1973; Felsenstein 1981).

6.6.2 Phylogenetics can reveal the origin of an emerging infectious diseases

In July 1981, the New York Times reported on rare cancer in 41 homosexual men in New York and California, 8 of whom had already died. This disease was named GRID for gay-related immunodeficiency disease.

In March 1982, the Washington Post reported about the same disease, now called Acquired Immunodeficiency Syndrome (AIDS), highlighting that it not only affects gay people. More than 1100 Americans have been already diagnosed with this disease, of which more than 400 have died. The disease was spreading rapidly, with more than 200 diagnoses made within the month prior to the report. Half of the victims were younger than 35 years old.

In 1983, a virus was isolated from patients with AIDS in two separate laboratories (Barré-Sinoussi et al. 1983; Gallo et al. 1983) and it was hypothesised that this virus, named Human

Immunodeficiency Virus (HIV), was actually the cause of AIDS (Marx 1984). It is now well-established that HIV causes AIDS (The Durban Declaration 2000). However, people and governments have long denied that HIV is the cause of AIDS; for example, the South African president still denied this causation in 2000 (The Durban Declaration 2000).

In 2022, 0.7% (mean, confidence interval $[0.6\%, 0.8\%]$) of adults in the age 15 to 49 years were estimated to be infected with HIV. The African Region remains the most severely affected region worldwide (World Health Organization 2023). Thus, it is essential that we obtain a detailed understanding of HIV dynamics so that appropriate actions can be taken to fight the epidemic. In the remainder of this section, we will discuss how the origin of the HIV epidemic was determined using phylogenetic trees. In Sections 6.6.3 and 6.6.4, we will discuss how we can investigate the spread of HIV using phylogenetic trees. As we will see, epidemic spread may be investigated to obtain public health knowledge or provide evidence in criminal cases.

6.6.2.1 HIV phylogeny reveals the origin of the epidemic

HIV must have evolved from some ancestor. However, no virus similar to HIV was known within the human population. Therefore, scientists started searching for viruses similar to HIV that infect species closely related to humans. The idea is that a virus infecting a closely related species (see Figure 6.1) might easily adapt and infect humans as well (zoonotic transmission).

A similar virus, found in most simian species, is the simian immunodeficiency virus (SIV). Many simian species are natural hosts to SIV, meaning SIV occurs in these species, often without causing disease. A zoonotic transmission from simians to humans was suggested.

A huge effort was made in the 1980s-1990s to collect SIV sequences from various simian species. Based on these sequences, scientists could reconstruct the maximum likelihood SIV/HIV phylogenies (Hahn et al. (2000, Figures 1 and 3)). Figure 6.25 shows such a tree. The tree highlights that HIV, in fact, comprises two genetically different groups: HIV-1 and HIV-2. Furthermore, it highlights that HIV-1 clusters with the chimpanzee (CPZ) virus and HIV-2 clusters with the sooty mangabee (SMM) virus. Phylogenies can further indicate the direction of transmission: for example, a tree indicates the direction from animals to humans if, when more and more sequences are added, the HIV sequences form a few nested clades within a large tree of SIV sequences. This is indeed what we observe when adding more sequences that were collected since the publication of the original trees.

Initially, the direction of transmission was not clear based on the phylogenetic tree. However, non-phylogenetic evidence, such as the fact that simians are a natural reservoir for SIV, their host range covers the areas where HIV appeared first, as well as results of evolutionary sequence analyses of SIV and HIV sequences, indicated that all HIV clades are results of zoonoses from simians to humans (Sharp, Robertson and Hahn 1995; Gao et al. 1999; Hahn et al. 2000).

In summary, the data suggest that HIV-1 is a zoonosis from chimpanzees, meaning that HIV-1 jumped from chimpanzees to humans (and analogously for HIV-2 and sooty mangabeys).

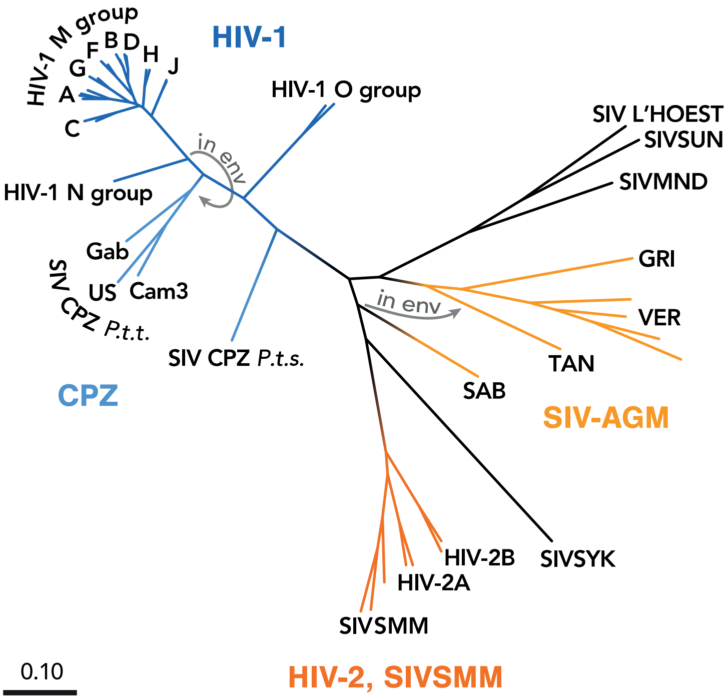


Figure 6.25: Phylogenetic tree on HIV and SIV samples indicating several zoonotic events.

Furthermore, the number of nested HIV clades within the SIV sequences is an estimate for the number of observed zoonoses (three for HIV-1, four for HIV-2, Hahn et al. (2000, Figure 3)).

6.6.2.2 How did HIV jump from simians to humans?

Understanding how HIV jumped from simians to humans requires considerations beyond phylogenetic analysis. Two main hypotheses were put forward to understand the jump.

Hypothesis 1 (polio hypothesis) Humans caused the HIV epidemic through contaminated polio vaccines in the 1950s in Africa. This hypothesis was popularised in the book "The river" by the journalist Edward Hooper (1999). W. D. Hamilton, one of the leading evolutionary biologists in the 20th century (kin selection, altruism; Hamilton's rule), set out to assess this hypothesis. Tragically, he died from malaria after returning from an expedition to Congo, aiming to find evidence for this hypothesis. Polio vaccines found in freezers later did not show any sign of contamination, disproving this hypothesis.

Hypothesis 2 (hunter hypothesis) HIV jumped from simians to humans in the course of hunting simians (Sharp et al. 2001). Since hunters may experience blood-to-blood contact with the hunted simians, such as through cuts, this could provide a potential transmission route. The hunter hypothesis is supported by the fact that areas with high SIV prevalence coincide with early HIV outbreaks, and hunting of simians occurs in these areas. Furthermore, the hunter hypothesis explains the observation that there has not been one but rather several introductions of HIV into the human population. Overall, evidence supporting the hunter hypothesis is very strong (Pépin 2021).

Today, the hunter hypothesis is the commonly accepted hypothesis for the origin of HIV.

6.6.3 The HIV epidemic in Switzerland

Within an epidemic, pathogen phylogenies can display transmission across different population groups. Kouyos et al. (2010, Figure 1) reconstructed a maximum likelihood phylogenetic tree on 5 700 HIV sequences from the Swiss epidemic, together with the same number of non-Swiss sequences. The Swiss sequences were collected from patients to screen for drug-resistant HIV strains in order to define a proper course of antiretroviral treatment. Each tip of this tree corresponds to a single consensus pathogen genetic sequence from an infected host (see Section 3.3.1). The colour of a tip and its pendant branch indicates the transmission group of the host associated with the tip: blue — Swiss intravenous drug users, red — Swiss men who have sex with men, cyan — Swiss heterosexuals, black — non-Swiss.

Clades of predominantly non-black tips nested within a subtree of black tips indicate the import of HIV from abroad into Switzerland. We call these clades “Swiss clades”. Swiss clades consisting of tips with more than one colour indicate HIV transmission between different transmission groups within Switzerland. Swiss clades of only one colour indicate that the disease is mostly spreading within a single Swiss transmission group rather than between groups. We observe that the red sequences often form small clusters within the black “backbone”. This indicates ongoing transmissions in the group of Swiss men who have sex with men. In contrast, the Swiss heterosexuals (cyan) and intravenous drug users (blue) are well-mixed within Swiss clades. This indicates that there is frequent transmission between the two transmission groups.

6.6.4 Phylogenetics in the court

6.6.4.1 HIV criminal case: Louisiana 1994

The following discusses the first court case in which phylogenetic methods were used as forensic evidence (Metzker et al. 2002). In 1994, a woman from Louisiana (USA) accused her ex-partner, a physician, of having infected her with HIV. While she tested HIV-positive, none of the ten men she reported sexual contact with during the previous decade were HIV-positive.

The woman claimed that the ex-partner purposefully infected her through an injection, which he claimed was a vitamin B boost, administered during a late-night visit to his practice. This visit to the physician's practice followed a fight between the victim and her ex-partner. The victim reported that her ex-partner did not want her to leave him, and if she did, he wanted to make sure she did not have sexual contact with anyone else anymore.

HIV only survives a couple of hours *in vitro*. However, it was observed that the physician took a blood sample earlier that day from an HIV-positive patient but never sent the blood to the laboratory.

How can one assess the victim's claim of intentional HIV transmission on the night of the "vitamin B boost"? For the first time in a criminal trial, phylogenetic methods were used to assess the claim. HIV samples from the victim, the suspected donor (the physician's patient), and 32 other HIV-infected individuals from the local area were obtained and sequenced. Based on these sequences, HIV phylogenies were reconstructed using various methods for two different genes. In every single reconstruction, the victim's sequences clustered within the suspected donor's sequences, and the remaining 32 sequences were further apart with bootstrap support (assessing the robustness of results; see Section 7.4.3) of $96\% - 100\%$, indicating that the most likely route of transmission was indeed from the suspected donor to the victim (see Metzker et al. (2002, Figures 1 and 2)).

Together with non-phylogenetic evidence, the physician was found guilty of having infected the woman with HIV. Additionally, he was found guilty of having infected the woman with Hepatitis C through the "vitamin B boost". As a result of these findings, he was sentenced to 50 years in prison.

6.6.4.2 Florida dentist

A dentist who died of HIV was suspected of having infected patients with HIV in the course of his practice (Ou et al. 1992). In a study by the Center for Disease Control and Prevention (CDC) published in 1992, viral samples of seven patients were sequenced, and a phylogeny was reconstructed. The analysis revealed that for five of the patients, the sequences clustered with the dentist's sequences and, together with epidemiological information, indicated the dentist to be the source of these infections.

This case induced a discussion on hygiene rules for healthcare workers. It remains unclear whether the dentist always wore protective gear during his practice (already in the 1980s, dentists were supposed to wear protective gloves). Generally, this was an exceptional case because in no other lawsuit involving HIV did a healthcare worker transmit the virus to a patient. To this day, it is not known if the transmissions by Florida's dentist occurred by chance or on purpose.

6.6.4.3 Bulgarian nurses in Libya

The last case we present occurred in Libya. Five Bulgarian nurses and a Palestinian doctor who were helping in Libyan hospitals were accused of having transmitted HIV to more than 400 children. They were sentenced to death by the Libyan government and were incarcerated for over eight years. During this time, data were collected, and computational methods were developed to show that the transmission occurred before the healthcare workers arrived in the country (de Oliveira et al. (2006)). Upon political pressure, the healthcare workers were finally freed in 2007.

7 Statistical testing

This chapter will explore how to determine which stochastic models to use in phylogenetic inference and how to summarise uncertainty in parameter estimates. At the core, we will test the plausibility of different hypotheses. A *null hypothesis* is the hypothesis that the data evolved under the null model, where the *null model* is some mathematical model (e.g. JC69 may be the null model; in Chapter 4, Section 4.1 the null hypothesis was that a variant has no effect on a disease). Testing the null hypothesis means testing whether the null model describes the data well. If not, we reject the null hypothesis. If the null hypothesis was tested against an alternative hypothesis, rejecting the null hypothesis means that we favour the alternative hypothesis, formalised as an alternative mathematical model. For example, think of the six-sided die. The null hypothesis could be that the die is fair, and rolling a 6 has the probability of 1/6. The alternative hypothesis in this case could be that the probability to roll a 6 is any value different from 1/6. We can also perform *model selection* without formulating the null hypothesis but by comparing different models using a statistical criterion.

Throughout this section, we always specify a mathematical model together with the hypotheses, so we will primarily use the word “model” to refer both to the model and the hypothesis. In what follows, we first discuss how to test the plausibility of a null model (or null hypothesis) $\mathcal{H}_0$ (Sections 7.1 and 7.2). Second, we will show how to perform model selection between several models $\mathcal{H}_0, \mathcal{H}_1, \ldots$ (Section 7.3). Then, given a model, we assess the uncertainty in the parameter and tree estimates (Section 7.4).

7.1 Test whether to reject the null model $\mathcal{H}_0$

In Chapter 4, we tested whether our data (a contingency table) rejected the null hypothesis that particular SNPs were not associated with a disease status (GWAS). The general idea was to determine the distribution $P(X = x|\mathcal{H}_0)$, where the random variable X stands for one field in the contingency table. The null model assumes that this random variable is hypergeometrically distributed (Box 9 on page 76) with parameters defined by the fixed row and column sums of the contingency table. Thus, for any possible observed data x, we can evaluate $P(X = x|\mathcal{H}_0)$, and can directly calculate the p-value for a particular observation (Box 1 on page 24) based on the hypergeometric distribution. If this p-value is below a given rejection threshold, we reject the null model $\mathcal{H}_0$.

In the context of phylogenetic tree inference, X is the random MSA, and $\mathcal{H}_0$ is a particular tree with some rate matrix Q. Here, it is generally impossible to determine the distribution

$P(X|\mathcal{H}_0)$ for the particular tree and Q as it would mean to evaluate the probability for all MSAs. However, there are roughly 4^{mn} MSAs for n sequences of length m (roughly as we did not consider the gaps). We thus proceed with a different type of test, as described in the next section.

7.2 Test whether to reject the null model $\mathcal{H}_0$ in favour of $\mathcal{H}_1$

We will compare $\mathcal{H}_0$ to other models. In general, a statistical test is defined through a *test statistic*, which is a function that transforms the data into real numbers, depending on $\mathcal{H}_0$ and $\mathcal{H}_1$. Depending on the value of the test statistics and the chosen significance level, we might reject the null hypothesis.

We will consider here likelihood-based test statistics (see Box 25 on page 116 for likelihoods in general and Section 6.3.3 for likelihoods in phylogenetics). We will first introduce the *likelihood ratio test (LRT)* to test if a null model $\mathcal{H}_0$ should be rejected when tested against model $\mathcal{H}_1$. Afterwards, we introduce the *Akaike Information Criterion (AIC)* which ranks different models $\mathcal{H}_0, \mathcal{H}_1, \mathcal{H}_2, \ldots$ according to their support based on likelihoods (Section 7.3).

7.2.1 Likelihood ratio tests (LRT)

Log-likelihood ratios can be used to determine confidence intervals for maximum likelihood estimators (see Box 26 on page 117) and, as we discuss here, also for model comparison. *Likelihood ratio testing (LRT)* allows us to test a null model $\mathcal{H}_0$ against a model $\mathcal{H}_1$. To use LRT, the null model needs to be a *nested model*, which means that the parameters of $\mathcal{H}_0$ are a subset of the parameters of $\mathcal{H}_1$. For example, JC69 is nested in K80 because when the two parameters of K80 have the same value, the model is reduced to JC69. We determine whether $\mathcal{H}_0$ should be rejected when tested against $\mathcal{H}_1$. We define the log-likelihood ratio function between the two models $\mathcal{H}_0$ and $\mathcal{H}_1$ on a given dataset as

$$\mathrm{LR}(\mathcal{H}_1, \mathcal{H}_0) = 2\log\left(\frac{L_1(\hat{\theta}_1)}{L_0(\hat{\theta}_0)}\right) = 2(\log L_1(\hat{\theta}_1) - \log L_0(\hat{\theta}_0)), \tag{7.1}$$

where L_1 and L_0 are the likelihood functions of models $\mathcal{H}_1$ and $\mathcal{H}_0$, and $\hat{\theta}_1$ and $\hat{\theta}_0$ are the maximum likelihood estimates under $\mathcal{H}_1$ and $\mathcal{H}_0$, respectively.

The likelihood value $L_1(\hat{\theta}_1)$ of the data under the general model will be higher or equal to the likelihood under the null model $L_0(\hat{\theta}_0)$. This is because the models are nested, and fixing some parameters of $\mathcal{H}_1$ to certain values will reduce it to model $\mathcal{H}_0$. Thus, it is always possible for $\mathcal{H}_1$ to obtain at least the same likelihood as $\mathcal{H}_0$. More formally, $2(\log L_1(\hat{\theta}_1) - \log L_0(\hat{\theta}_0)) \geq 0$. However, a positive difference does not necessarily mean that $\mathcal{H}_1$ is a better choice, as $\mathcal{H}_1$

requires more parameters. Overly general models tend to have reduced explanatory power, a phenomenon known as *overfitting*.

Now, we determine how big the log-likelihood ratio should be to reject $\mathcal{H}_0$ against $\mathcal{H}_1$. Assuming that $\mathcal{H}_0$ indeed generated the data, Wilk's theorem (Wilks 1938) states that if the sample size goes to infinity, the log-likelihood ratio is distributed as the χ^2 distribution:

$$\mathrm{LR}(\mathcal{H}_1, \mathcal{H}_0) = 2(\log L_1(\hat{\theta}_1) - \log L_0(\hat{\theta}_0)) \sim \chi^2_{df}, \tag{7.2}$$

where df is the degrees of freedom of the χ^2 distribution (Box 11 on page 78). df is calculated as the difference between the number of free parameters in $\mathcal{H}_1$ and the number of free parameters in $\mathcal{H}_0$. If some parameters of $\mathcal{H}_0$ are at the parameter boundary of $\mathcal{H}_1$ (e.g. 0 or ∞), they typically only count for 0.5 degrees of freedom (for more details see e.g. Self and Liang (1987)).

In a likelihood ratio test (LRT), we choose a level of significance α (see also Box 1 on page 24), and evaluate $\mathrm{LR}(\mathcal{H}_1, \mathcal{H}_0)$. We will reject $\mathcal{H}_0$ if and only if the log-likelihood ratio falls in the α-tail of the χ^2_{df} distribution, that is, if

$$p_{\chi^2_{df}}(x > \mathrm{LR}(\mathcal{H}_1, \mathcal{H}_0)) < \alpha. \tag{7.3}$$

The significance level α corresponds to the probability of falsely rejecting $\mathcal{H}_0$. We note that the mean and variance of the χ^2_{df} distribution increase linearly with df (see Box 11 on page 78), and so the threshold for rejecting the null model increases for increasing df.

Comparing our definition of LRTs and confidence regions (Box 26 on page 117) reveals that a null model is rejected at the level α if and only if its maximum likelihood parameters are not within the $(1 - \alpha)$-confidence region around the maximum likelihood estimates of $\mathcal{H}_1$. For more details, see also Section 7.4.1.

Example: rolling a die We again use the die rolling experiment as introduced in Box 25 on page 116 to illustrate the concept of likelihood ratio tests. We test whether we reject the null model $\mathcal{H}_0$ of a fair die when tested against the alternative model $\mathcal{H}_1$ of a loaded die. As we did above, we are only considering the probability of rolling a 6. In model $\mathcal{H}_0$, this probability is known and equal to $^1/_6$, so $\mathcal{H}_0$ has no free parameters (and the likelihood L_0 is constant for a given dataset). In model $\mathcal{H}_1$, this probability is equal to θ_1, the parameter of the model. We saw in Box 25 on page 116 that if we roll the die n times and obtain a six k times, the maximum likelihood estimate for the probability of rolling a 6 under model $\mathcal{H}_1$ is $\hat{\theta}_1 = {}^k/_n$ and that $L_1(\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$. The probability of rolling k sixes out of n rolls under the null model is $L_0 = \binom{n}{k}(\frac{1}{6})^k(\frac{5}{6})^{n-k}$. We can then calculate $\mathrm{LR}(\mathcal{H}_1, \mathcal{H}_0) = 2(\log L_1(\hat{\theta}_1 = {}^k/_n) - \log L_0)$, and check whether the value is in the α-tail of the χ^2_1 distribution. For $\alpha = 0.05$, the value is in the α-tail if $LR(\mathcal{H}_1, \mathcal{H}_0) > 3.84$. Thus, 3.84 is the rejection threshold. If the value is in the α-tail, we reject the null model $\mathcal{H}_0$ of our die being fair. Otherwise, we do not reject $\mathcal{H}_0$.

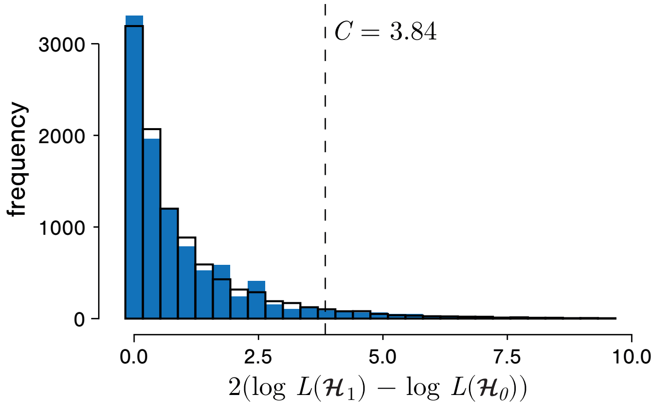


Figure 7.1: Histogram of log-likelihood ratio values obtained from $10\,000$ experiments with a fair die (blue) and the null hypothesis $\mathcal{H}_0$ being that the die is fair. In each experiment, the die is rolled $1\,000$ times, and k, the number of throws resulting in a 6, is recorded. The log-likelihood ratio is calculated for each value k and plotted as a histogram in blue. The χ^2_1 distribution is shown in black, and the boundary of the 0.05-tail of the distribution is indicated with the vertical dashed line.

For the die rolling experiment performed in Box 25 on page 116, where the die was rolled $n = 100$ times and a six was obtained $k = 40$ times, we calculate $\mathrm{LR}(\mathcal{H}_1, \mathcal{H}_0) = 30.62$, which is greater than 3.84. Thus, the null model is rejected at the $\alpha = 0.05$ level. Recall that we reject $\mathcal{H}_0$ precisely when the confidence interval around the maximum likelihood parameter estimate assuming $\mathcal{H}_1$ does not contain 1/6. Thus, equivalently, we can conclude from the interval calculated in Box 26 on page 117 that the null model is rejected for this experiment.

We also use the die rolling experiment to illustrate that the log-likelihood ratio function is well-approximated by a χ^2 distribution. We roll a fair die $n = 1\,000$ times and record k, the number of throws resulting in a 6. We repeat this experiment 10 000 times. Then we calculate the log-likelihood ratio for all 10 000 experiments and plot the histogram for $\mathrm{LR}(\mathcal{H}_1, \mathcal{H}_0)$ (Figure 7.1, blue). The experimental histogram corresponds very well to the $\chi^2_{df=1}$ distribution, as we can see in Figure 7.1, black. In this figure, we also plot the 95-th percentile of the χ^2 distribution (the boundary of the 0.05-tail; dashed line), which is the value C such that $p_{\chi^2_{df=1}}(x > C) = 0.05$ (thus only 5% of the distribution fall to the right of this blue line). Here $df = 1$, thus we have $C = 3.84$.

Example: phylogenetics For a fixed tree and given MSA, we can calculate the maximum likelihood estimates for different substitution models, for example, the JC69 and the K80 model. We then determine the log-likelihood ratio and reject JC69 at the 0.05 level if the log-likelihood ratio is bigger than 3.84, since $df = 1$ (see the overview of model parameters of nucleotide substitution models in Table 5.1).

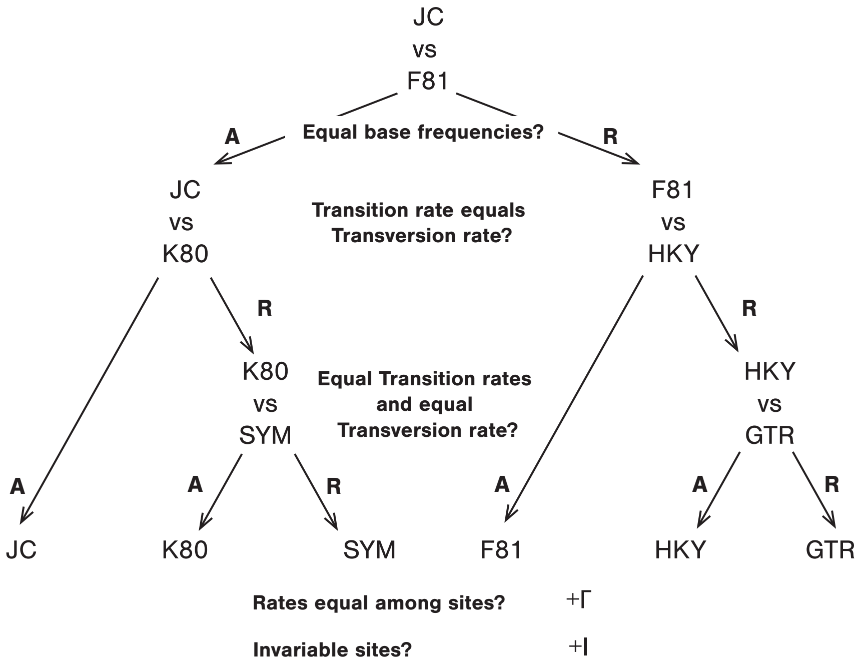


Figure 7.2: Decision tree showing the successive LRTs performed by the software ModelTest (Posada and Crandall 1998) to identify an appropriate substitution model. The first test is shown at the top, and the possible outcome of each test is shown as A (accept $\mathcal{H}_0$) or R (reject $\mathcal{H}_0$).

The LRT only considers two models; however, in many phylogenetic questions, there are many more candidate models to select from, such as, for example, JC69, K80, HKY and GTR. If we want to choose an appropriate substitution model for a given MSA on a fixed tree, we may need to perform several successive LRTs. An example of a scheme for such model comparison is shown in Figure 7.2. One caveat of such a model selection scheme is that it involves multiple tests. Correcting for multiple tests can be done easily only if the number of tests is known in advance. This is, however, not the case here. Another caveat with this specific example is that some models may not be tested at all, depending on the previous tests. For instance, if the very first test (JC69 vs F81 (see Sections 5.3.1 and 5.3.3)) rejects $\mathcal{H}_0$, then we will proceed directly to testing F81 against HKY and will not even consider the K80.

Summary of the likelihood ratio tests Likelihood ratio tests are a statistical test for which the test statistic is the log-likelihood ratio. The user decides on the level of significance as introduced in Box 1 on page 24 ($\alpha = 0.05$ is often used), which allows the rejection threshold of the test to be determined. For LRTs with 1 degree of freedom at $\alpha = 0.05$, the rejection threshold is 3.84 (compare also to the rejection threshold calculation for multiple testing in

	$\mathcal{H}_0$ true	$\mathcal{H}_0$ false
Reject $\mathcal{H}_0$	Type I error	Correct
Accept $\mathcal{H}_0$	Correct	Type II error

Table 7.1: Two type of errors in statistical tests.

Section 4.1.3). If the calculated test statistic is greater than the rejection threshold, we will reject $\mathcal{H}_0$ for the given data.

Rejecting a null model $\mathcal{H}_0$ does not imply that the general model $\mathcal{H}_1$ is a good model to explain the data. In fact, when testing $\mathcal{H}_0$ against $\mathcal{H}_1$, we can obtain very small $p-$values simply because $\mathcal{H}_0$ is a very bad null model, but not because $\mathcal{H}_1$ is a good model. Addressing the overall fit of a model has to be done with procedures such as, for example, described in Section 7.1. Alternatively, we can explore a wider range of models (Section 7.3).

7.2.2 Errors in statistical testing

When performing statistical tests (e.g. the LRT), it is important to distinguish between two types of errors, shown in Table 7.1. A *type I error* occurs if the $\mathcal{H}_0$ model is true, but we falsely reject it. A *type II error* occurs if $\mathcal{H}_1$ is true, but we fail to reject $\mathcal{H}_0$. The *accuracy* and *power* of a test depend on type I and type II errors, respectively. Accuracy is defined as $(1 - (\text{type I error}))$ and power as $(1 - (\text{type II error}))$. There is a trade-off between the two, as increasing accuracy will decrease power and vice-versa.

The significance level chosen by the user is equal to the type I error, meaning the user directly controls the accuracy of a test. Evaluating a statistical test's power often requires simulations under $\mathcal{H}_1$. In general, the power increases with the difference between model $\mathcal{H}_1$ and the null model $\mathcal{H}_0$.

Example: rolling a die We illustrate statistical errors, accuracy, and power with the die-rolling experiment. These results are based on our particular realisation of the experiment and will slightly deviate for each experiment, as the outcome of each experiment is random.

First, we experiment with a fair die. This experiment will determine the accuracy and the type I error. We perform 10 000 experiments, each consisting of 1 000 die rolls, and record the number of times we roll a 6 (as above). We find that $\mathcal{H}_0$ is rejected in 5.1% of the experiments. This simply confirms that the significance level we have chosen is 0.05, and thus the accuracy is 0.95.

Next, we assess the test's power and type II error. We assume the die is unfair, meaning the null model is wrong. First, we assume $\theta = 1/5$. Performing the same experiments as before, we end up correctly rejecting $\mathcal{H}_0$ in 78% of the experiments. Thus, we estimate that the power is 0.78 and the type II error is 0.22.

Lastly, we consider another unfair die, for which $\theta = 1/2$. In all 10 000 experiments we correctly reject $\mathcal{H}_0$. The power is, therefore, 1, and the type II error is 0.

7.3 Compare models $\mathcal{H}_0, \mathcal{H}_1, \mathcal{H}_2, \ldots$ using the Akaike information criterion

As seen in the previous section, the likelihood ratio test can only be used for two models where one is nested within the other. To compare more models that are not nested, we can use the *Akaike information criterion (AIC)* (Akaike 1974). The AIC ranks models according to their fit to the data.

The AIC of a particular model i is:

$$\mathrm{AIC}_i = -2 \log L_i(\hat{\theta}_i) + 2p_i, \tag{7.4}$$

where p_i is the number of free parameters of the model, L_i is the likelihood function under the model, and $\hat{\theta}_i$ are the maximum likelihood parameter estimates of the model.

The AIC needs to be calculated separately for each model we want to compare. The model with the lowest AIC is the model which fits the data best, according to the criterion. The AIC is related to the expected Kullback-Leibler divergence (MacKay 2003); that is, AIC aims to measure the loss of information compared to the true (unknown) model. The $+2p_i$ term has the effect of penalising models with more parameters, which helps combat overfitting.

It is important to note that the absolute AIC values are not informative on their own. Only the difference between the AIC values is informative. The difference represents the difference in loss of information compared to the (unknown) true model. The model with the minimum AIC value is picked. Again, we do not know how well this best model explains the data overall; we only know that it explains the data better than any of the other considered models. Models with an AIC within $1-2$ of the minimum also have substantial support. Models with an AIC within about $4-7$ of the minimum have considerably less support, while models with an AIC more than 10 above the minimum have essentially no support (Burnham and Anderson 2002, page 71).

Example: substitution model selection The AIC has been used in many studies to perform substitution model selection. As an example, we use a maximum likelihood tree topology of the *rbcL* genes from 12 plant species shown in Yang (2014, Figure 4.16), reconstructed with

Model	p	l	AIC
JC69	21	$-6\,262.01$	$12\,566.02$
K80	22	$-6\,113.86$	$12\,271.72$
HKY	25	$-6\,101.76$	$12\,253.52$
JC69+Γ_5	22	$-5\,937.80$	$11\,919.60$
K80+Γ_5	23	$-5\,775.40$	$11\,596.80$
HKY+Γ_5	26	$-5\,764.26$	$11\,580.52$

Table 7.2: Number of parameters p, maximum log-likelihood values l, and the AIC value for different substitution models. Model parameters and branch lengths were optimised by Yang (2014) on a fixed topology of the 12 species in Yang (2014, Figure 4.16).

an HKY+Γ_5 model[1]. Based on that tree topology, the author estimated the maximum likelihood substitution model parameters and branch lengths for different models of nucleotide substitution. We use these results with permission of the author and state the maximum log-likelihood values l for the different models in Table 7.2.

In what follows, we employ the AIC to determine which substitution model fits this tree topology best. The number of parameters p for each model is calculated as follows: there are $2n - 3$ branches in a tree with n tips (Section 6.2.3.1), so for the tree of 12 species, there are $2 \times 12 - 3 = 21$ free parameters (the branch lengths), in addition to the parameters of the substitution model minus one. The reason for removing one free parameter is that the substitution rate matrix and the branch lengths are correlated for trees where all tips are sampled at the same time: multiplying all rates by a factor k and dividing all branch lengths by the same factor k will give the same log-likelihood value (refer to Sections 5.3.5 and 6.4.3). Thus, we can fix the average substitution rate to 1, meaning that, for example, for the JC69 model, the number of parameters is $p = 21 + 1 - 1 = 21$.

The HKY+Γ_5 has the lowest AIC value, making it the best-suited model for this dataset. The next best model is K80+Γ_5. The difference in AIC values between K80+Γ_5 and HKY+Γ_5 is 16.28, meaning that the simpler K80+Γ_5 has essentially no support. The likelihood ratio tests in Yang (2014) also support the HKY model with rate heterogeneity the most.

In this example, the tree topology was fixed. The AIC can also be used to compare models if the tree topology is not fixed. This contrasts with the LRT, where the topology needs to be fixed. If the topologies are different, we have an additional model or model parameter for the

[1] The Γ_5 refers to a model with rate variation among sites as described in Section 5.5, using 5 discrete rate categories to approximate the continuous Γ distribution for computational tractability.

tree topology. If the two topologies are different, the two models or parameters are different, and thus, the models are not nested, meaning we cannot apply LRT and need to use AIC.

We emphasise that on a fixed topology, we can assume two different nested substitution models and $n-1$ branch length parameters, and for each substitution model, estimate its maximum likelihood parameters together with the maximum likelihood branch lengths. Since the models are nested, we can apply the LRT, meaning in the example above, we could also perform LRTs (as done in Yang (2014)).

Substitution model selection using LRT and AIC has been automated in software tools such as jModelTest (https://evomics.org/learning/phylogenetics/jmodeltest/) (Posada 2008; Darriba et al. 2012), which can test dozens of models on a fixed tree or co-estimate trees using PhyML (PhyML (http://www.atgc-montpellier.fr/phyml/) (Guindon and Gascuel 2003)).

7.4 Assessing uncertainty in estimates

After selecting the best model for a given dataset, we may also want to estimate the uncertainty associated with the maximum likelihood parameter estimates under this model. We want to know how wide the range of “likely” values for each parameter is. This range is called a confidence interval for a single parameter or *confidence region* for one or more parameters. Confidence intervals are formally defined in Box 26 on page 117. In what follows, we provide four procedures for obtaining estimates for the confidence intervals. We note that for many biological problems, we cannot freely choose between the four procedures; rather, we may be limited to one particular approach, depending on the models and the datasets.

We illustrate the confidence intervals again using the die-rolling example. We perform one die roll experiment consisting of 1 000 die rolls and record the number of times we roll a 6. We recorded rolling six 165 times, meaning the maximum likelihood estimate $\hat{\theta}$ for the probability of rolling a six is 0.165.

7.4.1 Obtaining confidence intervals using the LRT

In Box 26 on page 117, we described how to construct confidence intervals for maximum likelihood parameter estimates. The $(1-\alpha) \times 100\%$ confidence interval is the set of parameter values θ for which the estimated $\hat{\theta}$ (under a $\mathcal{H}_1$) would not lead to rejection of the hypothesis that θ ($\mathcal{H}_0$) equals the true parameter at the significance level α. Thus, likelihood-based confidence intervals are directly related to likelihood ratio tests. We show the confidence interval based on an LRT for the die rolls example in Figure 7.3.

Most of our substitution models contain several parameters; therefore, we obtain a confidence region composed of the intervals for all the parameters. This region is easy to characterise if it is, for example, a circle in the two-dimensional case with the maximum likelihood estimate

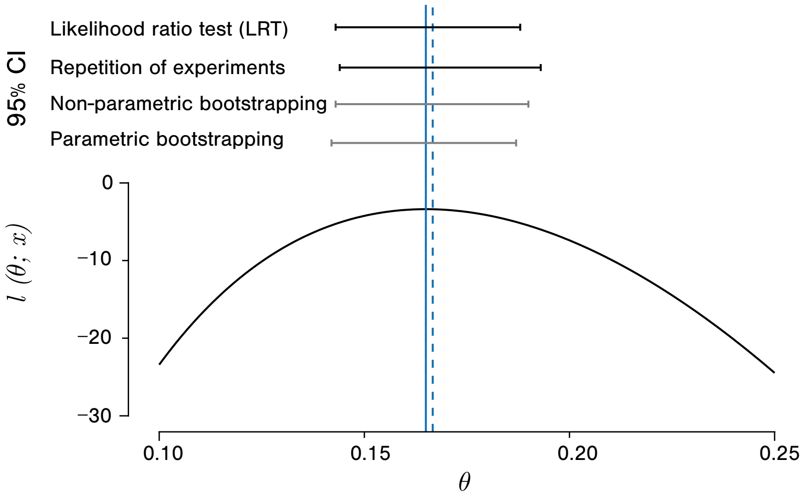


Figure 7.3: Confidence interval estimates. A fair die is rolled $1\,000$ times, and a six was recorded 165 times. From this experiment, we estimate the probability of rolling a six, θ. The black curve is the log-likelihood curve for θ. The maximum likelihood estimate for θ is shown in blue (solid line), and the true probability of $^1/_6$ is shown next to it (blue dashed line). The 95% confidence intervals obtained using the LRT, repetition of experiments, non-parametric bootstrapping, and parametric bootstrapping are shown as horizontal lines.

in the centre. However, often it is very different from a circle. To facilitate confidence calculations, *profile likelihoods* are considered (e.g. see Cole, Chu and Greenland (2014)). To calculate the profile likelihood, we fix all parameters except one to their maximum likelihood estimates. The profile confidence interval for the un-fixed parameter is then calculated in the same way as in the LRT-based confidence interval for the one-parameter case. From this, it follows that in the special case when we have only one parameter, the profile confidence interval equals the classic confidence interval as introduced in Box 26 on page 117. For example, in the die experiment, we have only one parameter; thus, the profile confidence interval is the same as the confidence interval.

LRT-based confidence intervals (or profile confidence intervals in case of many parameters) are straightforward to calculate for parameter estimates of a given substitution model when a fixed phylogenetic tree is provided. In Section 5.4.3, we calculated LRT-based confidence intervals for pairwise distances between sequences under the JC69 model.

7.4.2 Obtaining confidence intervals by redoing experiments

If we have access to the experimental system used to produce the dataset, we can repeat the experiment multiple times. Each experiment will produce a different maximum likelihood estimate for the parameter and, based on the definition of confidence intervals in Box 26 on page 117, the 95% confidence interval of the parameter can be obtained by discarding the bottom 2.5% and the top 2.5% estimates. This procedure can, for instance, be used with our example of a die, with a bacterial evolution experiment, and so on. For the die experiment, Figure 7.3 shows the 95% confidence interval obtained via 100 repetitions of the experiment.

The LRT-based and redoing experiment-based 95% confidence intervals are not identical in general, as shown, for example, in Figure 7.3. This is because these confidence intervals are different approximations to the true confidence interval. The LRTs rely on the χ^2 approximation. Redoing experiments leads to an approximation of the CI if a finite number of experiments is performed (only infinite repetition leads to an exact one).

In most biological applications, redoing experiments to obtain the confidence interval is not feasible. However, we introduce this framework here as it leads to the exact confidence intervals (if enough replicates are generated), and the non-parametric bootstrapping approach in the next section approximates the procedure of redoing experiments.

7.4.3 Obtaining confidence intervals by non-parametric bootstrapping

In general, it is impossible to rerun real-life experiments such as the evolution of mammals. However, we can mimic the repetition of experiments through *bootstrapping*. This term stems from the English idiom, “to pull yourself up by your own bootstraps,” which alludes to creating something from nothing. In statistics, this refers to a process of generating “new” datasets by somehow shuffling/sampling from an existing dataset.

Bootstrapping may be done within a *parametric* or *non-parametric* framework. In the parametric framework, an explicit model with a finite number of parameters is defined for generating the data. In the non-parametric framework, such an explicit model assumption is avoided. Thus, under a non-parametric framework, a model is either completely avoided or the model structure, such as the number of model parameters, grows with the available data (see Section 9.2.3).

In non-parametric bootstrap, the procedure consists of randomly sampling our dataset with replacement until we get a second dataset of the same size as the original. It is important that the sampling is carried out with replacement (meaning we put the sample back into the dataset before performing another sampling step); otherwise, we simply recover the original dataset. Resampling the dataset multiple times and obtaining the maximum likelihood estimate of the parameters for each bootstrap dataset will allow us to construct a 95% confidence interval by ignoring the 2.5% smallest and largest maximum likelihood estimates.

For our die rolling experiment, the 95% non-parametric bootstrapping confidence interval is obtained by considering our initial experiment where we rolled six 165 times out of 1 000 tries. To obtain a bootstrap dataset, we sample another 1 000 die rolling results by sampling from the 1 000 results of the initial experiment with replacement. We then obtain the maximum likelihood estimate for the probability of rolling a six for this bootstrap dataset. In Figure 7.3, we show the 95% non-parametric bootstrapping confidence interval. If the original dataset was big enough, it is straightforward to see that this procedure leads to a confidence interval indistinguishable from the one we would obtain by redoing experiments (previous section). Note that if the dataset is small, bootstrapping underestimates the size of the true confidence interval.

In Section 7.4.5, we will use non-parametric bootstrapping to obtain confidence measures on tree topologies.

7.4.4 Obtaining confidence intervals by parametric bootstrapping

Finally, we can estimate confidence intervals by parametric bootstrapping. Here, additional datasets are simulated under the model $\mathcal{H}_1$ with parameters $\hat{\theta}$, the maximum likelihood estimates for the dataset. For these parametric bootstrap datasets, we estimate the maximum likelihood parameters again, and the 95% parametric bootstrapping confidence interval is obtained, as above, by ignoring the 2.5% smallest and largest maximum likelihood estimates.

For our die-rolling experiment, we obtained 100 parametric bootstrap datasets by performing our experiment with a die where the probability of rolling a six is 0.165. The result is shown in Figure 7.3. For our die, the confidence interval based on parametric bootstrapping would be equivalent to the confidence interval based on redoing experiments if the maximum likelihood estimate equals the true parameter.

Parametric bootstrapping will be used for phylodynamic parameter inference in Section 9.1.6.3.

7.4.5 Tree uncertainty estimation

Estimating the uncertainty in reconstructed trees is a unique challenge, as trees are complex objects with a discrete component, namely the topology. In particular, a continuous uncertainty interval or region is not defined for phylogenies. Instead, a set of trees collectively represents the uncertainty in trees; this set can then be summarised further.

The LRT method cannot be used to assess uncertainty as comparing different topologies corresponds to non-nested models (see Section 7.3). Furthermore, we typically cannot rerun evolution and obtain additional trees by redoing experiments. Obtaining parametric bootstrapping confidence estimates is only possible if we have a model for tree growth (see Chapter 9). Non-parametric bootstrapping, however, provides a useful tool for estimating the uncertainty in a tree.

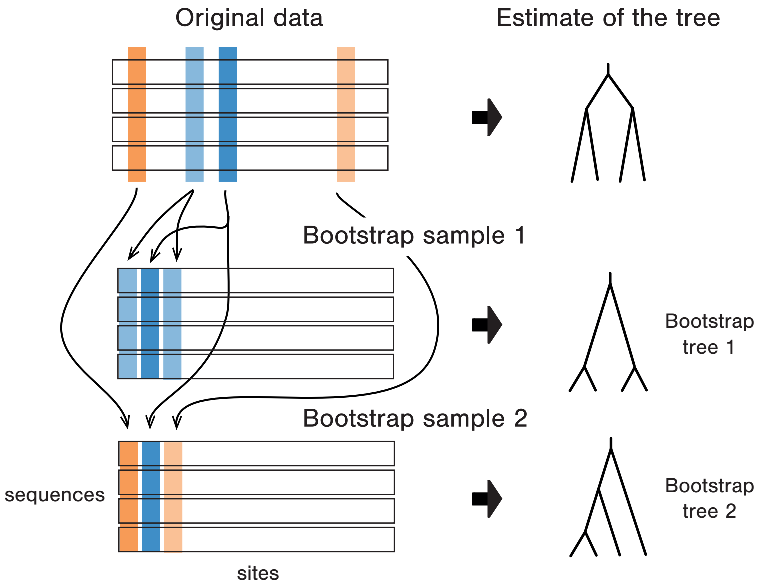


Figure 7.4: Bootstrap for phylogenies. Consider an MSA with m columns. A bootstrap MSA is obtained by sampling m columns with replacement from the original MSA. If m is very large, the bootstrap samples show the same statistical properties as if obtaining additional samples by repeating evolution (which, of course, is not feasible in practice). Figure inspired by Felsenstein (2003).

The procedure (first introduced in Felsenstein (1985a)) is as follows: we obtain a bootstrap MSA by sampling m sites (columns) at random with replacement from the original MSA of m sites, as shown in Figure 7.4. Note that the order of the rows representing different samples is kept unchanged during the bootstrap procedure. If the original MSA was long enough, the resulting bootstrap MSAs show approximately the same variation as would be obtained from repeating the evolution of those organisms multiple times. This becomes clear when considering die rolling: if we roll the die 1 000 times (which corresponds to repeating evolution) or if we roll it 1 000 times and then select 1 000 outcomes with replacement (which corresponds to bootstrapping), we obtain roughly the same number of outcomes of 6.

We then reconstruct a maximum likelihood bootstrap tree from each newly created bootstrap MSA. The set of bootstrap phylogenies can be summarised by comparing it to the original maximum likelihood tree. A common comparison criterion is the following: for each node in the original maximum likelihood tree, we count how many bootstrap trees have an internal node whose set of descending tips is the same as for the node in the maximum likelihood tree. In other words, for each clade in the original maximum likelihood tree, we count how many bootstrap trees contain this particular clade. For each node in the maximum likelihood tree,

Box 29: Maximum likelihood tree inference

Selecting a substitution model, reconstructing a phylogenetic tree, estimating all model parameters for a multiple sequence alignment (MSA), and providing uncertainty estimates can be done as follows under the maximum likelihood framework:

1. for each plausible substitution model, infer a maximum likelihood tree and choose the tree with the topology and the branch lengths that maximise the likelihood;
2. determine the tree and substitution model with the highest support using AIC and proceed with that tree and model in steps 3 and 4;
3. determine the confidence intervals for the model parameters using the LRT;
4. determine the confidence in the maximum likelihood tree by non-parametric bootstrapping of the MSA.

we divide this count by the number of bootstrap trees. The resulting values, which are between 0 and 1, are called the *bootstrap supports* of the internal nodes in the original maximum likelihood tree. Figure 7.5 shows the result of bootstrapping on an SIV-HIV MSA within a maximum likelihood phylogenetic analysis; each node of the maximum likelihood tree is marked with its bootstrap support.

Taken together, the maximum likelihood framework for tree reconstruction introduced in Chapter 6 with the concepts on statistical testing and uncertainty introduced in this section allow for a coherent tree reconstruction approach acknowledging uncertainty, as summarised in Box 29 on page 198.

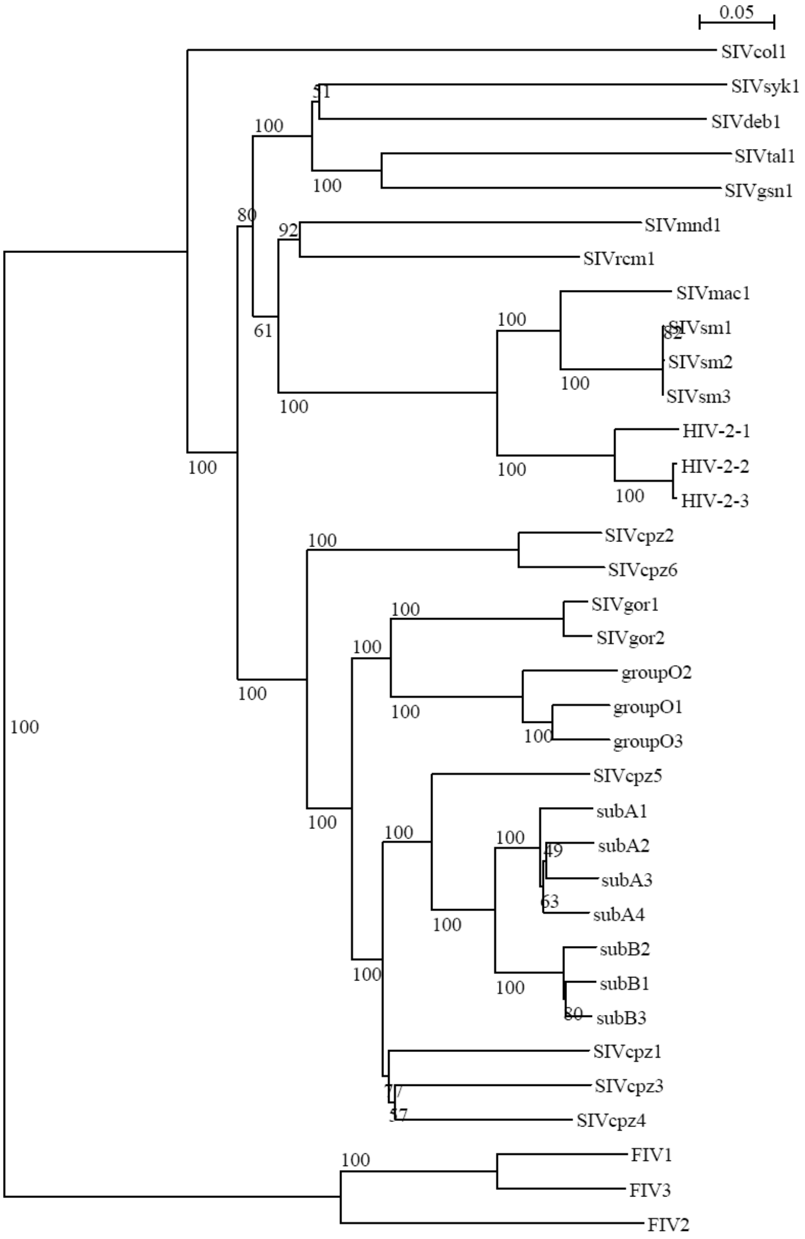


Figure 7.5: Maximum likelihood tree of SIV-HIV sequences with bootstrap support (between 0 and 100) marked for each internal node. Figure adapted from Sägesser (2010).

8 Traits and comparative methods

The distribution of phenotypic traits among individuals or species is the result of evolution. These traits are encoded in *characters*. Characters can be *discrete*, such as spike numbers of HIV virions, leg numbers of arthropods, or the presence of a fur pattern in rodents. Characters can also be *continuous*, such as height, weight, virulence, or dinosaur jaw length.

Comparative methods are a large family of approaches that study trait evolution, often with the goal of learning whether the evolution of specific characters is correlated. The earliest of these methods ignored potential phylogenetic relationships between samples. However, in the late 1970s, it was recognised that ignoring phylogenetic dependencies can lead to highly incorrect conclusions (see Felsenstein (2003) for the historical perspective). For this reason, trait evolution is now generally studied in the context of phylogeny.

In this chapter, we will consider first discrete and then continuous trait evolution and will talk about *association* and *correlation* between characters. Two characters are associated when knowledge of one character informs us about the second character, meaning the characters are not independent, while correlation of two characters is a special association where the characters follow a particular trend such as a linear relationship (Altman and Krzywinski 2015).

We will demonstrate how phylogenetic relationships between sampled individuals or species can easily lead to apparent associations between pairs of independently evolving characters when phylogenetic relationships are ignored during the analysis. We will then introduce phylogenetic comparative methods that explicitly account for phylogeny, allowing evolutionary correlations to be detected.

Note that throughout this chapter, we will assume, for the sake of simplicity, that the phylogenetic tree is known. In reality, we rarely know the true tree. Thus — when applied to empirical data — the methods presented here are generally combined with phylogenetic inference procedures of the kind discussed in Chapters 6 and 10.

8.1 Assessing associations between discrete characters

Suppose we want to find out whether there is an association between eye colour and feather colour in a group of eight bird species shown in Figure 8.1. While the sample shows only two combinations (yellow eyes and green feathers, red eyes and blue feathers), the other two combinations (yellow eyes and blue feathers, or red eyes and green feathers) might also be

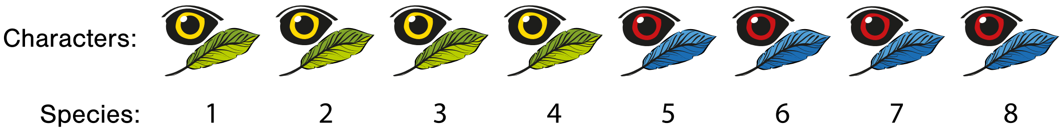


Figure 8.1: The combinations of discrete characters — eye and feather colour — for eight sampled species. Eyes in this example are either yellow or red, feathers green or blue.

possible, even though they are not present in our data. From looking at the data, it is tempting to conclude that there is a strong evolutionary correlation between the two characters; that is, if a species evolves to have green feathers, this will likely coincide with the appearance of yellow eyes.

8.1.1 Assuming independence across samples

One way to quantify the significance of this correlation would be to employ Fisher's exact test (see Box 10 on page 77). Recall that we need to formulate the null hypothesis before performing this statistical test. The null hypothesis in our case is:

$\mathcal{H}_0$: *Having yellow eyes is equally likely among green- and blue-feathered species.*

For Fisher's exact test, we fill out a contingency table with our observations (see Table 8.1), and we can then formulate the null hypothesis mathematically:

$\mathcal{H}_0$: *The number of species with both yellow eyes and green feathers is drawn from a hypergeometric distribution, given the total number of species with each of the four individual trait values (the row and column sums of the contingency table).*

The hypergeometric distribution describes an urn experiment. In this case, we can use the urn analogy to represent the number of green balls obtained when drawing 4 balls without replacement (representing the individuals with yellow eyes) from an urn with 4 green balls (representing green-feathered individuals) and 4 blue balls (representing the blue-feathered individuals). We denote the number of green balls among the drawn 4 balls with the random variable Z. We calculate the p-value of the observed outcome, a, which is the probability of obtaining the observed or more extreme outcomes. Here, we define more extreme as obtaining more species with yellow eyes and green feathers than observed. We set the significance level to 0.05. If the obtained p-value is less than the significance level, we reject the null hypothesis, as it is very unlikely that we could get such data under the null model.

In the above example, we observe four individuals with yellow eyes and green feathers. Let us now compute the probability of this event under hypothesis $\mathcal{H}_0$ by following Fisher's exact

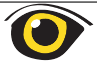
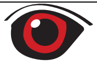
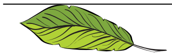
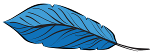
			Totals
	4	0	4
	0	4	4
Totals	4	4	8

Table 8.1: Contingency table for the feather and eye colours using the example in Figure 8.1. No species in the sample has green feathers and red eyes or blue feathers and yellow eyes; four species have green feathers and yellow eyes, and four species have blue feathers and red eyes.

test (Box 10 on page 77):

$$P(Z = 0|\mathcal{H}_0) = \frac{\binom{4}{4}\binom{4}{0}}{\binom{8}{4}} \simeq 0.0143. \tag{8.1}$$

As we cannot observe more green balls than 4 (a more extreme event is impossible), the p-value is 0.0143, well below a significance level of 0.05. We thus reject the null hypothesis of an equal distribution of yellow eyes among green- and blue-feathered species at the 0.05 significance level. Instead, there seems to be a significant correlation between the two features.

8.1.2 Considering phylogenetic relatedness

The analysis above relies on the assumption that the individuals in our samples are independent from one another. Let us now assume that, in reality, our eight species are related by the evolutionary history shown in Figure 8.2. The colours along the branches depict the history of the evolution of feather and eye colour traits.

When accounting for this relatedness, instead of considering the traits at the tips, we consider the changes along the branches of the phylogenetic tree. In particular, we test whether two characters change on the same branch more often than expected under a null model. Again, we summarise the example data in a contingency table shown in Table 8.2.

Suppose again we model our changes using an urn experiment. Let us use red balls to represent the branches where the feather colour trait underwent an evolutionary change (here, 1 red ball) and black balls to represent the branches this trait did not change (here, 13 black balls). Now, we draw k balls without replacement (here $k = 1$) to determine the branches on which the change of eye colour occurred. The number of red balls drawn follows the hypergeometric distribution (Box 9 on page 76).

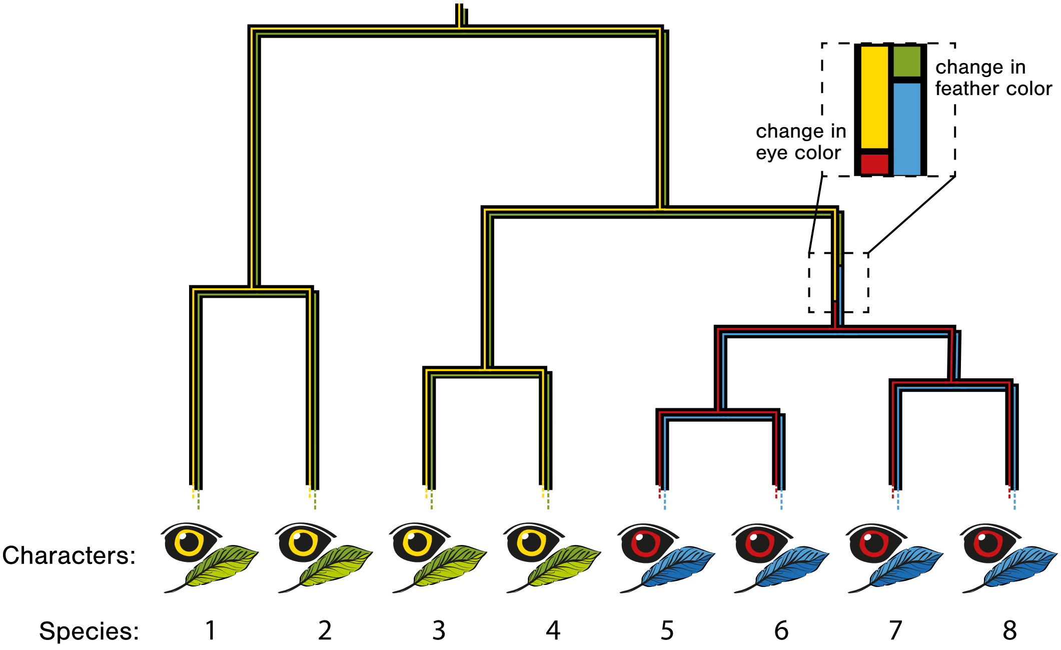


Figure 8.2: The phylogenetic tree connecting the eight sampled species from Figure 8.1. The species at the root of the tree has yellow eyes and green feathers, and so do four of its descendants. In this particular example, the changes in feather and eye colours occurred only once, and both took place along the same branch.

From this new perspective, our null hypothesis is:

$\mathcal{H}_0$: *The number of branches with a change in both feather and eye colour follows a hypergeometric distribution.*

The probability of the data under the null hypothesis is:

$$P(\text{both feather and eye colour change on a branch}|\mathcal{H}_0) = \frac{\binom{1}{1}\binom{13}{0}}{\binom{14}{1}} \simeq 0.0714. \qquad (8.2)$$

The p-value is obtained by summing over the probability of the data and the probability of any more extreme result. More extreme corresponds to more than one branch having both a feather and eye colour change. As we only have one change of each trait, the p-value equals 0.0714. As the p-value is higher than a significance threshold 0.05, we do not reject the null hypothesis that changes are equally likely on every branch and thus are not correlated at the 0.05 significance level.

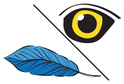
	Change	No change	Totals
Change	1	0	1
No change	0	13	13
Totals	1	13	14

Table 8.2: The contingency table for changes per branch based on the tree in Figure 8.2. In this example, there are no branches with a single change, one with changes in both characters and 13 with no changes.

This is an entirely different result from the one we arrived at when we did not consider the phylogeny. It shows that to detect correlated evolution between traits, we must take into account the phylogenetic relationships between individuals or species.

8.1.3 Detecting associations between discrete variables in empirical data

While the above analysis demonstrates the necessity of considering phylogeny when testing for correlated evolution, it does not immediately lead to an obvious practical means of detecting such correlations. The reason is that for real datasets, we typically know neither the phylogeny nor the pattern of trait evolution that proceeds down it (that is, we do not have the full picture of the phylogeny with character changes along branches like the one shown in Figure 8.2).

While we will not delve into the details here, we note that several studies have proposed coherent methodology to test for correlated evolution in discrete characters, starting with Ridley (1983) who proposed a test using the parsimony algorithm (see Section 6.3.2) to reconstruct the tree and the ancestral changes. Later, it was explicitly recognised that longer branches have a higher chance of co-occurring changes than shorter branches, which is taken into account in analyses using likelihood-based methods (Pagel 1994). For an extended discussion on this topic, refer to Felsenstein (2003, Chapter 25).

8.2 Assessing associations between continuous characters

Just as with discrete characters, spurious correlations between continuous characters can be induced by shared ancestry when ignoring this ancestry and, consequently, common evolution during analysis. The only difference is how these characters evolve through time. Evolution occurs due to a continuous time jump process for discrete characters, with instantaneous transitions (changes) occurring between the allowed states. For continuous characters, evolution occurs according to a process that can take any state in a continuous space. In both cases, evolution is such that points on the tree that are close together (less time passed) are more likely to share the same trait value than points that are far apart (more time passed). This means that groups of closely related individuals are more likely to share similar trait values than distantly related individuals. Failing to account for this can lead to incorrect conclusions since the similarity of additional samples from the closely related group can be incorrectly interpreted as evidence for correlation.

Consider the distribution of a pair of continuous traits — for example, wing span and beak length — corresponding to the same set of species as discussed in the previous section, illustrated in Figure 8.3. The individual values are the result of a random process, which we detail later in this chapter. The qualitative effect of the phylogeny is already clear. For instance, the wing span trait values (blue) for species 1 and 2 are almost identical, as are the corresponding beak length trait values (black) for the same species. Similarly, all of the blue values for the remaining six species have similar values, as do the remaining black values. In other words, although we have eight distinct species in this dataset, the number of independent species samples of each trait value is less. It may be as low as two since we observe two main clades. Looking only at the trait values in the absence of the phylogeny, the presence of these closely related species in the dataset can lead to the conclusion that the evolution of the two traits is correlated, even when no such correlation exists in the underlying random process.

To quantify this effect, we need a quantitative test for correlation. In the following, we describe a quantitative test for correlation ignoring the phylogeny (analogously to Section 8.1.1 for the discrete case). In Section 8.2.3, we describe a quantitative test for correlation given a phylogeny (analogously to Section 8.1.2 for the discrete case).

8.2.1 Assuming independence across individuals

Imagine that two distinct traits X and Y (wing span and beak length in our example) are recorded for each of n sampled species, with the numerical trait values for species i given by x_i and y_i.

A common approach to detecting simple correlation between continuous variables is *linear regression*. In this method, we assume that the observations (the trait values) are related to each other linearly such that

$$y_i = \beta x_i + b + \epsilon_i, \tag{8.3}$$

where β and b are unknown constants describing the slope and offset of the relationship. The terms ϵ_i are “error” or “noise” terms, which account for any deviation from the linear relationship. The error terms are assumed to be:

1. statistically independent from one another, and

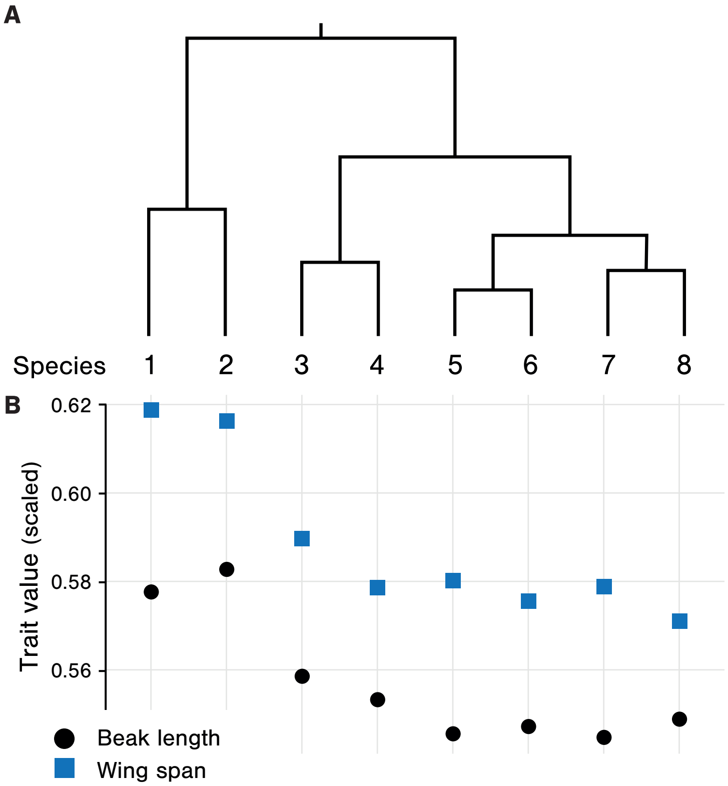


Figure 8.3: A The phylogeny relating eight bird species and **B** continuous wing span and beak length traits of the eight species. The trait values highly depend on which of the two main clades of the phylogeny a given species belongs to.

2. drawn from the same normal distribution Normal$(0, \sigma^2)$ (see Box 13 on page 80) for all data points.

This means that the error terms are independent and identically distributed.

Fitting this linear model to the data requires we find estimates $\hat{\beta}$ and $\hat{b}$ that minimise the difference between the measured values y_i and the predicted values $f_i = \hat{\beta}x_i + \hat{b}$. Once the model is fit to the data, we can compute the *coefficient of determination* R^2. It quantifies the amount of variance in Y that is explained by X. Given the mean of the observed data $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, R^2 is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(f_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{8.4}$$

Therefore, a R^2 value close to 1 means that the variation in Y is perfectly explained by X, while a value close to 0 indicates that X is unable to predict the variation in Y through linear regression.

Performing these calculations for the trait data shown in Figure 8.3 **B** yields an R^2 value of 0.98, indicating a very strong correlation between the two traits. The p-value computed from the corresponding t-statistic[1] is 2.49×10^{-6}, suggesting that this correlation is highly significant.

However, we know that these individuals have a shared evolutionary history, shown in the phylogenetic tree in Figure 8.3 **A**. The fact that traits evolved on this phylogeny means that their noise terms are neither independent nor identically distributed, meaning that the assumptions of a classic linear regression are not met. Thus, applying classic linear regression directly to the two sets of trait values is not a reliable means of testing for the correlated evolution of the characters since the assumed null model is incorrect.

The following section will define a null model for continuous trait evolution along a phylogeny. Then, we will show how to use linear regression in combination with this improved null model to detect significant correlations between characters more reliably.

8.2.2 Modelling continuous trait evolution with Brownian motion

A simple model to describe the evolution of continuous traits along a phylogenetic tree is the *Brownian motion model* (Box 30 on page 209). Figure 8.4 illustrates how Brownian motion can be applied to continuous trait evolution on a tree. In Figure 8.4 **A** we see a phylogeny of four extant species, 1-4. Figure 8.4 **B** shows one possible evolution of a single continuous character under the Brownian motion model on this phylogeny. Individuals 1, 2, and 3 share a common evolutionary history; their characters evolve together from the root until the branching in the phylogenetic tree (dashed lines in Figure 8.4 **B**).

If we assume Brownian motion for each trait as the null model, the trait values of 1, 2, and 3 are not independent. Thus, we cannot perform standard linear regression to investigate correlations between these traits evolving along the tree.

8.2.3 Considering phylogenetic relatedness using the contrast method

One method that lets us account for the interdependencies of evolutionary traits on trees — using Brownian motion as the null model — is the *contrast method* developed by Joseph Felsenstein (1985b). In the contrast method, we mathematically eliminate interdependence in our data by performing linear regression not on the original data points but on recomputed

[1] The t-statistic is defined as $\hat{t} = {}^{\hat{\beta}}/_{s_{\hat{\beta}}}$, where $s_{\hat{\beta}}$ is the standard error of $\hat{\beta}$. The t-statistic obeys a well-known distribution and can be used to test for statistical significance, giving a p-value under the null hypothesis that no correlation existed ($\beta = 0$) (Fisher 1925).

Box 30: Brownian Motion

Brownian motion is named after the Scottish botanist Robert Brown, who formulated it after observing crop seed movement in water under a microscope. He noticed that the crop seeds move randomly on the water surface due to hits from water molecules. Another example of this type of movement is a ball thrown around by a crowd in a football stadium. Different people push the ball around, resulting in a random movement pattern.

Brownian motion is the motion resulting from a Wiener process, which is a continuous time stochastic process. Note that the terms Wiener process and Brownian motion process refer to the same process. The Wiener process is named after the American mathematician Norbert Wiener (1894-1964). The Wiener process is defined as a stochastic process $(W_t)_{t \in T}$ and $T \subseteq \mathbb{R}$ that fulfils the following four conditions:

1. $W_0 = 0$: the process starts in 0;
2. W_t is almost surely continuous: $P(W_t \text{ continuous}) = 1$;
3. W_t has independent increments which implies memorylessness: for $0 \leq s_1 \leq t_1 < s_2 \leq t_2$, $(W_{t_1} - W_{s_1})$ and $(W_{t_2} - W_{s_2})$ are independent;
4. for $0 \leq s \leq t$, the difference $W_t - W_s$ is distributed as a normal distribution with variance that depends on the time difference ($W_t - W_s \sim \text{Normal}(0, \sigma^2(t-s))$).

Note that thus far, we considered a memoryless stochastic process on a discrete state space (Markov processes in Box 24 on page 98), while the Wiener process is a memoryless stochastic process on a continuous state space. The Wiener process is important for many fields, including statistical physics and quantitative finance. As such, detailed information about the process can be found in most books introducing stochastic processes (see e.g. Bertoin (1994)).

values that are independent and identically distributed. The recomputed values are referred to as *contrasts*. We will illustrate the contrast method on the example phylogeny on four tips in Figure 8.5.

In the following, we assume to have $j = 1, \ldots, m$ distinct continuous characters, each evolving down the same phylogenetic tree in Figure 8.5, starting at the root, according to the Brownian motion model with variance parameter σ_j^2. We denote the value of the observed trait $j \in [1, m]$ at node $k \in [1, 4]$ by X_k^j. We use the same notation for the unobserved trait values at the internal nodes where $k > 4$. Finally, the branch length leading to node k is t_k.

It is clear from the tree that the trait values X_1^j and X_2^j are not independent observations, as they share evolutionary history through branches t_6 and t_5. In what follows, we use the trait values at the tips and the branch lengths to define a set of independent variables — *normalised contrasts* — such that they possess identically normally distributed errors under the assumption that the traits evolved according to the Brownian motion model. With the normalised contrasts, we can perform linear regression to test for additional correlation between the traits besides that due to the common ancestry.

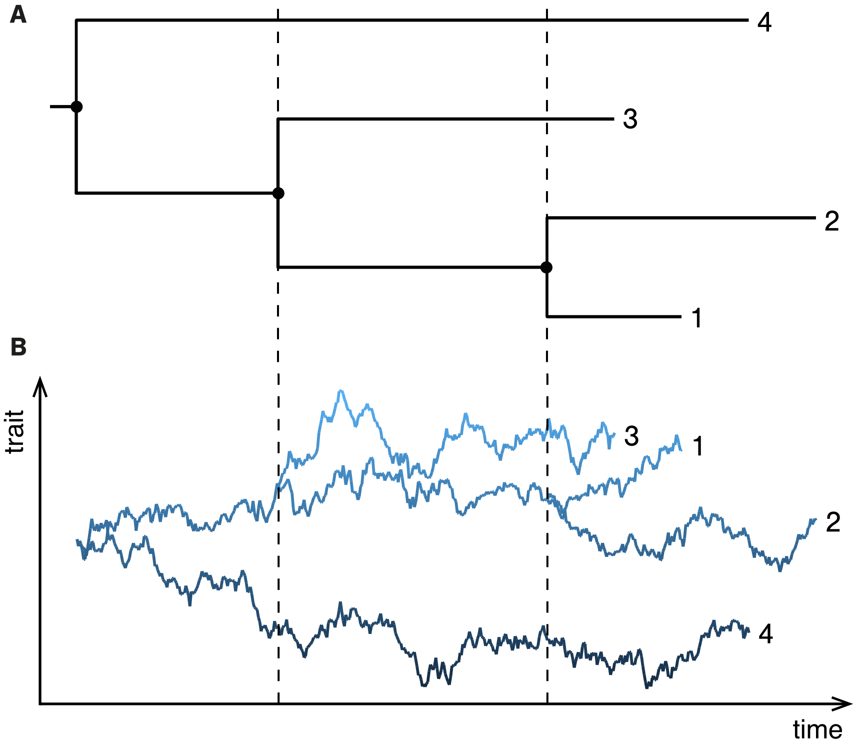


Figure 8.4: **A** A phylogeny of four individuals and **B** a realisation of trait evolution along the phylogeny according to the Brownian motion model. When a branching event happens in the phylogeny, the stochastic process modelling trait evolution splits into two separate and independent stochastic processes, starting with the trait value of the ancestral individual upon branching.

8.2.3.1 Preliminaries

Before proceeding, we will first prove two lemmas that will be useful in the following calculations.

In what follows, the following general identities for random variables will be used repeatedly:

$$\begin{aligned}\operatorname{Var}(\alpha U+\beta V) &= \alpha^2 \operatorname{Var}(U)+\beta^2 \operatorname{Var}(V)+2\alpha\beta \operatorname{Cov}(U,V), &(8.5)\\ \operatorname{Cov}(Y,\alpha U+\beta V) &= \alpha \operatorname{Cov}(Y,U)+\beta \operatorname{Cov}(Y,V), &(8.6)\\ \operatorname{Cov}(U,U) &= \operatorname{Var}(U), &(8.7)\end{aligned}$$

where U, V and Y are arbitrary real random variables, and α and β are real constant values

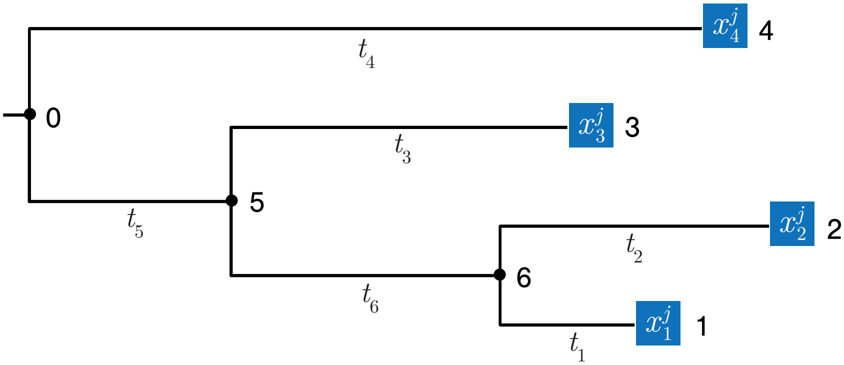


Figure 8.5: A phylogenetic tree on four tips on which the contrast method is illustrated. The extant nodes are labelled $1, \ldots, 4$, the root 0, and the internal nodes 5, 6. The length of the branch leading to node k is denoted with t_k. The observed value of trait j in node k is denoted by x_k^j (blue boxes).

(positive or negative); variance and covariance were defined in Section 1.3.

Lemma 8.2.1. *The variance of the trait variable at a leaf a is determined by the sum of the branch lengths between the leaf and the root, t_{sum}, as*

$$\mathrm{Var}(X_a^j) = \sigma_j^2 t_{sum}. \tag{8.8}$$

Proof. Firstly, note that the random variable that represents the trait value at a leaf node (node 1, for example) can be expanded as a sum of the following differences:

$$X_1^j = (X_1^j - X_6^j) + (X_6^j - X_5^j) + (X_5^j - X_0^j) + X_0^j. \tag{8.9}$$

Since X_0^j is fixed at 0 (Brownian motion property 1) and due to the independence of increments (Brownian motion property 3), we have

$$\mathrm{Var}(X_1^j) = \mathrm{Var}(X_1^j - X_6^j) + \mathrm{Var}(X_6^j - X_5^j) + \mathrm{Var}(X_5^j - 0). \tag{8.10}$$

Furthermore, due to the linear dependence of the variance of the Brownian motion process on time (Brownian motion property 4), we have

$$\mathrm{Var}(X_1^j) = \sigma_j^2 t_1 + \sigma_j^2 t_6 + \sigma_j^2 t_5 = \sigma_j^2 (t_1 + t_6 + t_5), \tag{8.11}$$

meaning the variance of the trait variable at the leaf is determined by the sum of the branch lengths between the leaf and the root. While we used a particular example tree and node, this proof works for any tree and node. □

Lemma 8.2.2. *The covariance between the random variables corresponding to the trait values at any two nodes a and b is determined by the sum of the lengths of the ancestral branches they share, t_{sum}, as*

$$\mathrm{Cov}(X_1^a, X_2^b) = \sigma_j^2 t_{sum}. \tag{8.12}$$

Proof. Consider the covariance between the trait variables at two different tip nodes, for example, tips 1 and 2. To do this, note that, like X_1^j, X_2^j can also be expanded as a sum of changes along branches:

$$X_2^j = (X_2^j - X_6^j) + (X_6^j - X_5^j) + (X_5^j - 0). \tag{8.13}$$

To compute the covariance between X_1^j and X_2^j, we successively apply the covariance identities and invoke Brownian motion properties 3 and 4 to find

$$\mathrm{Cov}(X_1^j, X_2^j) = \sigma_j^2(t_6 + t_5). \tag{8.14}$$

Thus, the covariance between the random variables corresponding to the trait values at any two nodes is determined by the sum of the lengths of the ancestral branches they share. Again, while we used a particular tree and node pair here, this proof works for any tree and node pair. □

8.2.3.2 Contrast method

The key idea behind the contrast method is to focus on contrasts between the trait values observed for species pairs rather than the absolute trait values. These contrasts depend only on the evolution that has occurred along the branches between the two species and their most recent common ancestor, but not on the evolution along the parts of the phylogeny that the two species share (the branches from the most recent common ancestor of the two species to the root). Our goal is to transform the observed trait values into another set of values (namely the contrasts) that are mutually independent and identically distributed under the null model of Brownian motion.

The contrasts are defined through a recursive procedure. We now state the overall recursive idea, referring to the core equations derived in later parts.

In the recursion, repeat until no cherries are left in the tree:

1. identify a cherry in the phylogeny with child nodes c_1 and c_2 and parent node p;
2. compute a normalised contrast for each trait $\hat{Z}^j_{(c_1,c_2)}$ as in Equation (8.19); if this is the only cherry remaining, finish here;
3. prune the chosen cherry from the tree, replacing it with a new tip node p';
4. compute new trait values $\hat{X}_p^j$ for this node as in Equation (8.28) and the length $\hat{t}_p$ of the branch above it as in Equation (8.30) (or terminate if this node is the root).

By construction, the pruning step leaves us with a new tree with a set of trait values that are statistically independent of the previously calculated contrast. Thus, for each trait, a set of $n-1$ independent normalised contrasts are calculated for a tree on n individuals, while the normalisation ensures that the contrasts are identically distributed.

Unlike the raw trait values, the contrasts, under the null model of Brownian motion, satisfy the assumption of the linear regression correlation test introduced above (independence and identical distribution) and can be used to test for the presence of evolutionary correlations between continuous character, beyond those correlations naturally introduced by the shared phylogeny.

8.2.3.3 Contrast for a cherry

Consider the values of trait 1, observed at the cherry consisting of leaf nodes 1 and 2 in Figure 8.5. We define the contrast for a cherry as the difference in trait values:

$$Z^j_{(1,2)} = X^j_1 - X^j_2. \tag{8.15}$$

Due to the independence of increments property of the Brownian motion process (Box 30 on page 209), this contrast is statistically independent of the trait value at every other leaf on the tree.

To normalise this contrast to a standard deviation of 1, we need to compute the variance of $Z_{(1,2)}$. To do this, consider that

$$Z^j_{(1,2)} = (X^j_1 - X^j_6) + (X^j_6 - X^j_2). \tag{8.16}$$

Again, because of the independence of these two increments, we have

$$\mathrm{Var}(Z^j_{(1,2)}) = \mathrm{Var}((X^j_1 - X^j_6)) + \mathrm{Var}((X^j_6 - X^j_2)), \tag{8.17}$$

which, by the property 4 of the Brownian motion , becomes

$$\mathrm{Var}(Z^j_{(1,2)}) = \sigma^2_j t_1 + \sigma^2_j t_2. \tag{8.18}$$

Thus, the normalised contrast is

$$\hat{Z}^j_{(1,2)} = \frac{Z_{(1,2)}}{\sqrt{\mathrm{Var}(Z_{(1,2)})}} = \frac{X^j_1 - X^j_2}{\sigma_j\sqrt{t_1 + t_2}}. \tag{8.19}$$

8.2.3.4 Pruning the tree

As outlined in Section 8.2.3.2, we now prune the cherry corresponding to the contrast calculated in Section 8.2.3.3 and replace it with a new tip. Then, a new set of trait values needs to be associated with this new tip, such that a further cherry can be picked from the resulting tree to calculate another contrast. We now explain how to associate the new trait value.

The key requirement for the new trait value (in our example $\hat{X}_6^j$) is that it needs to be statistically independent of the cherry contrast $Z_{(1,2)}$. That is, we require that

$$\mathrm{Cov}(Z^j_{(1,2)}, \hat{X}^j_6) = 0. \tag{8.20}$$

Given this goal and that $Z^j_{(1,2)}$ is the difference between the trait values 1 and 2, it is natural to propose that an independent variable might be obtained via the weighted average of X_1^j and X_2^j. Thus, we assume

$$\hat{X}^j_6 = fX^j_1 + (1-f)X^j_2. \tag{8.21}$$

To determine which f fulfils the condition in Equation (8.20), we combine Equations (8.20) and (8.21):

$$\begin{aligned}
\mathrm{Cov}(Z^j_{(1,2)}, fX^j_1 + (1-f)X^j_2) &= f\,\mathrm{Cov}(Z^j_{(1,2)}, X^j_1) + (1-f)\,\mathrm{Cov}(Z^j_{(1,2)}, X^j_2)\\
&= f\,\mathrm{Cov}(X^j_1 - X^j_2, X^j_1) + (1-f)\,\mathrm{Cov}(X^j_1 - X^j_2, X^j_2)\\
&= f\,\mathrm{Cov}(X^j_1, X^j_1) - f\,\mathrm{Cov}(X^j_2, X^j_1)\\
&\quad + (1-f)\,\mathrm{Cov}(X^j_1, X^j_2) - (1-f)\,\mathrm{Cov}(X^j_2, X^j_2)\\
&= f\,\mathrm{Var}(X^j_1) - (1-f)\,\mathrm{Var}(X^j_2) + (1-2f)\,\mathrm{Cov}(X^j_1, X^j_2).
\end{aligned} \tag{8.22}$$

Using Lemmas 8.2.1 and 8.2.2 we can substitute

$$\mathrm{Var}(X^j_1) = \sigma^2_j(t_1 + t_6 + t_5), \tag{8.23}$$

$$\mathrm{Var}(X^j_2) = \sigma^2_j(t_2 + t_6 + t_5), \tag{8.24}$$

$$\mathrm{Cov}(X^j_1, X^j_2) = \sigma^2_j(t_6 + t_5), \tag{8.25}$$

to obtain

$$\begin{aligned}
\mathrm{Cov}(Z^j_{(1,2)}, fX^j_1 + (1-f)X^j_2) &= \sigma^2_j\big(f(t_1 + t_6 + t_5) - (1-f)(t_2 + t_6 + t + 5)\\
&\quad + (1-2f)(t_6 + t_5)\big)\\
&= \sigma^2_j\,(ft_1 + (1-f)t_2) = 0.
\end{aligned} \tag{8.26}$$

Solving for f, we find that Equation (8.20) is satisfied when

$$f = \frac{t_2}{t_1 + t_2}. \tag{8.27}$$

Substituting this back into Equation (8.21) yields the following expression for the new trait value:

$$\hat{X}_6^j = \frac{t_2 X_1^j + t_1 X_2^j}{t_1 + t_2}. \tag{8.28}$$

This implies that $\hat{X}_6^j$ is simply the average of the trait values at the descendant leaves when the branch lengths on either side of the cherry (here t_1 and t_2) are the same.

To determine further contrasts in the pruned tree using this new trait value, we also need to compute its variance and its covariance with other tips in the tree. Again, using the properties of the Brownian motion process on the tree, we find:

$$\begin{aligned}
\operatorname{Var}(\hat{X}_6^j) &= f^2 \operatorname{Var}(X_1^j) + (1-f)^2 \operatorname{Var}(X_2^j) + 2f(1-f) \operatorname{Cov}(X_1^j, X_2^j) \\
&= \sigma_j^2 \left({}^2(t_1 + t_6 + t_5) + (1-f)^2(t_2 + t_6 + t_5) + 2f(1-f)(t_6 + t_5)\right) \\
&= \sigma_j^2 \big(t_6 + t_5 + t_2 + f^2(t_1 + t_2 + 2t_6 + 2t_5) \\
&\quad - 2f(t_2 + t_6 + t_5) + 2f(t_6 + t_5) - 2f^2(t_6 + t_5)\big) \\
&= \sigma_j^2 \left(t_6 + t_5 + f^2 t_1 + (1-f)^2 t_2\right) \\
&= \sigma_j^2 \left(t_6 + t_5 + \frac{t_1 t_2}{t_1 + t_2}\right).
\end{aligned} \tag{8.29}$$

Note that the variance is almost the variance of the unknown true trait value X_6^j at node 6, $\sigma_j^2(t_6 + t_5)$, but with an additional increment. This is equivalent to the variance one would see if one were to extend the original branch length t_6 to

$$\hat{t}_6 = t_6 + \frac{t_1 t_2}{t_1 + t_2}. \tag{8.30}$$

It is straightforward to show that the covariance of the new trait value $\hat{X}_6^j$ with the trait value at any other (non-descendant) leaf is simply the variance contributed by the shared branches. For example, the covariance between $\hat{X}_6^j$ and X_3 is given by

$$\begin{aligned}
\operatorname{Cov}(\hat{X}_6^j, X_3) &= f \operatorname{Cov}(X_1, X_3) + \operatorname{Cov}(X_2, X_3) \\
&= f\sigma_j^2 t_5 + (1 - f\sigma_j^2 t_5) \\
&= \sigma_j^2 t_5.
\end{aligned} \tag{8.31}$$

With that, the pruning step becomes the following:

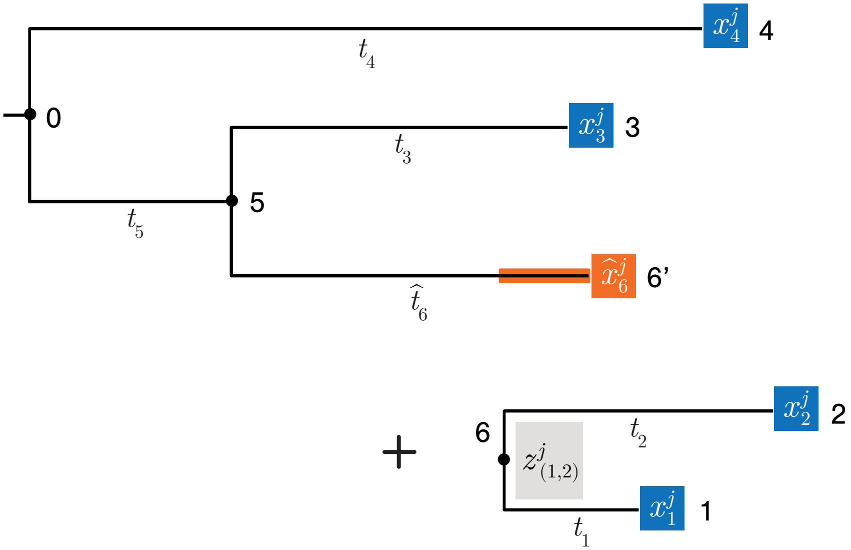


Figure 8.6: Pruning step of the contrast method. The cherry $(1, 2)$ for which the first normalised contrast was calculated in the original tree (Figure 8.5) is removed, and replaced with a new tip node $6'$ that has trait values defined by $\hat{X}_6^j = {}^{(t_2 X_1^j - t_1 X_2^j)}/_{(t_1+t_2)}$ and a parent branch of length $\hat{t}_6 = {}^{(t_6+t_1 t_2)}/_{(t_1+t_2)}$. A cherry in the pruned tree is chosen to calculate the next normalised contrast.

1. replace the cherry corresponding to the first contrast with a new leaf node $6'$;
2. define the trait values at this new leaf node using Equation (8.28);
3. lengthen the branch above the cherry according to $\hat{t}_6$ using Equation (8.30).

This procedure is summarised in Figure 8.6, which illustrates the decomposition of the original tree in Figure 8.5 into both the first normalised contrast $\hat{Z}^j_{(1,2)}$ and the pruned tree with the new leaf node $6'$, trait value $\hat{X}_6^j$ and branch length t_6'.

8.2.3.5 Some notes on the application of the method

In the above method, we made two key assumptions:

1. we assumed that the starting state of the Brownian motion process was the same (zero) for every character, and
2. we assumed that the variance accumulation parameter σ_j^2 of each trait-specific Brownian motion process was known.

In practice, neither of these assumptions is likely to be met. However, neither of them matters when we use linear regression to assess the independence of characters. For the first, this is simply because the contrasts depend only on differences between observed values. Thus, unlike the raw trait values, the contrasts for a given trait are always centered around zero — adding an offset due to the starting value at the root does not change this property.

To see why the second assumption does not matter, consider what would happen if the traits evolved under the Brownian motion processes with distinct trait-specific variance parameters, σ_j^2, but performing the above contrast calculations assuming a fixed variance accumulation parameter of 1. For a given trait, it is easy to see from Equation (8.19) that the normalised contrasts computed under this incorrect assumption would still be identically normally distributed with some variance, but this variance would no longer equal 1. Therefore, the set of normalised contrasts belonging to each trait would each be normally distributed with a variance unique to that trait.

In this situation, are the assumptions of the linear regression correlation test still met? The answer is yes. To see why, let us rewrite Equation (8.3) in terms of (realisations of) the normalised contrast variables $\hat{Z}^1_{(c_1,c_2)}$ and $\hat{Z}^2_{(c_1,c_2)}$. The linear regression assumption is then:

$$\hat{z}^1_{(c_1,c_2)} = \alpha\hat{z}^2_{(c_1,c_2)} + b + \epsilon_{(c_1,c_2)}. \tag{8.32}$$

All we require is that $\epsilon_{(c_1,c_2)}$ is drawn from a fixed normal distribution with a given variance and a mean of 0 for some value of b which captures constant offsets, and any value of α. That is, we require $\hat{Z}^1_{(c_1,c_2)} - \alpha\hat{Z}^2_{(c_1,c_2)}$ to be normally distributed around some mean and variance. Given that both $\hat{Z}^1_{(c_1,c_2)}$ and $\hat{Z}_{(c_1,c_2)}$ are normally distributed, this is immediately satisfied due to the properties of normally distributed random variables (Box 13 on page 80).

Thus, for the purposes of linear regression, contrasts can be computed assuming any fixed starting value for the Brownian motion process and any variance parameter value (1 is a natural choice).

8.2.4 Examples using the contrast method

Independent contrasts for the example tree with two clades As a concrete demonstration of the application of this method, we will consider again the 8-species, 2-trait example shown in Figure 8.3. We will not give detailed calculations here; we will only provide the results. By calculating the contrasts for each of the two characters and plotting the results (see Figure 8.7), we see that the previous evidence for correlation has essentially disappeared. In particular, the R^2 value between these two sets of contrasts is approximately 0.34, much weaker than between the raw trait values, for which the R^2 was 0.92. Similarly, the p-value for rejecting the null hypothesis (no correlation between the contrasts of different characters) is 0.168 compared to the p-value of 10^{-4} obtained from the raw trait values. Thus, the contrasts in this example offer no support for the hypothesis that there is a correlation. We highlight that there could still be a correlation, but we did not observe it due to a type II error.

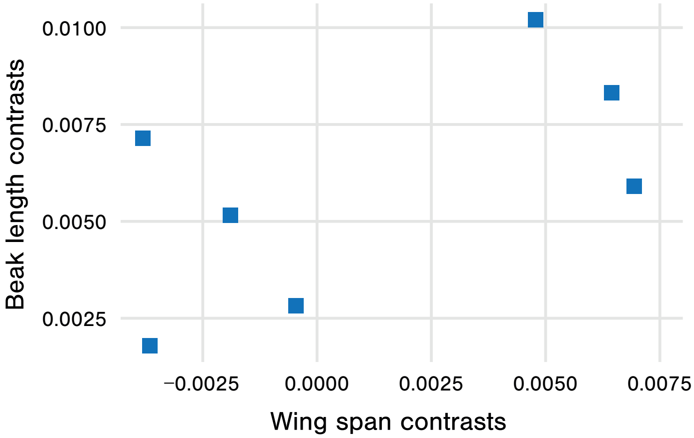


Figure 8.7: The contrasts computed for the pair of continuous characters shown in the example Figure 8.3. The contrasts, from which the effect of the shared ancestry has been removed, do not show evidence for correlated evolution. Recall that a simple regression that does not remove the effect of phylogenetic relatedness suggested a strong correlation.

Example: Evolution of bird nesting behaviour An organism's "life history" is a term used to encompass the many details of its life: how long it takes to mature, its life span, the number of offspring, and so on. Like any heritable biological traits, those involved in the life history can evolve and are potentially subject to selection. Martin (1995) presented a detailed study of how aspects of the life histories of birds vary across species. More specifically, the author studied a variety of life history traits of 123 species of North American birds from the orders *Passeriformes* and *Piciformes*, seeking to understand some of the main evolutionary and environmental influences behind the observed variation in these traits. Identifying whether or not there existed evidence for correlated evolution among pairs of life history traits was the principal goal of the study.

Among the studied continuous traits were body mass, nest predation (probability of nest failure due to predation), clutch size (average number of offspring in a nest), adult survival rates (probability of adult birds surviving over a fixed time period) and annual fecundity (offspring produced by female birds in a year). Table 8.3 lists the p-values obtained for three distinct pairs of these traits by applying linear regression first to the raw trait values and then to the independent contrasts. While all three pairs display significant correlation when studied using the uncorrected trait values, only adult survival vs. annual fecundity retains its statistical significance when phylogeny is taken into account. This means that the available data only provide evidence that this particular pair of traits may be evolving in a correlated way and further drives home the importance of accounting for shared ancestry when conducting such analyses.

Trait comparison	p-value (traits)	p-value (independent contrasts)
Predation vs. clutch size	< 0.001	0.089
Body mass vs. annual fecundity	< 0.001	0.052
Adult survival vs. annual fecundity	< 0.001	< 0.001

Table 8.3: Comparison between regression p-values obtained by Martin (1995) for pairs of avian life history traits, with (right-most column) and without (middle column) accounting for phylogenetic relatedness using the phylogenetic independent contrasts method.

8.3 Extensions

This chapter introduced basic modelling approaches to assess the association between characters. For continuous characters, we assumed a Brownian motion model of evolution, and for discrete characters, we assumed the hypergeometric distribution for changes along branches.

Many more sophisticated methods have been suggested that build upon these approaches (see e.g. Felsenstein (2003) and Garamszegi (2014)). In particular, the book by Luke Harmon (2018) provides a useful guide.

We highlight two extensions of the basic approaches. First, we only considered methods that compare discrete characters with discrete characters or continuous characters with continuous characters. More generally, however, it is useful to study evolutionary correlations between continuous and discrete characters. Such comparisons are the focus of Felsenstein (2012).

Second, we assumed a process without biological constraints by using Brownian motion for continuous and hypergeometric distribution for discrete characters. In particular, under the Brownian motion model, the variance in a trait increases linearly with time without any bounds. However, many traits (for example, the body size of terrestrial animals) are bounded. One way to address such constraints is to replace the basic Brownian motion process with another continuous stochastic process, namely the Ornstein-Uhlenbeck process, which is essentially a Brownian motion process with a forcing term that ensures a stationary state. This idea was initially proposed by Felsenstein (1988) and first used by Hansen (1997). Recent comparative methods such as Mitov, Bartoszek and Stadler (2019) allow for arbitrary combinations of Brownian motion and Ornstein-Uhlenbeck models.

8.4 Connections to GWAS

We now comment on parallels between this chapter and Chapter 4, where we considered genome-wide association studies (GWAS). Both here and in the GWAS chapter, we are interested in detecting correlations between characters. However, GWAS is classically applied only to single nucleotide polymorphisms (SNPs) in sexually reproducing organisms undergoing recombination. The SNP sites are generally separated far enough apart on the genome such that their linkage is broken up quickly through sexual reproduction. Thus, these SNPs can be considered completely unlinked, meaning that the phylogenetic relationships between the samples are different for every SNP. This assumption implies that each individual is an independent data point, and null models, such as the hypergeometric distribution of trait values (as used in Chapter 4), may be appropriate. Putting this independence of samples into a tree concept means that the relationships between all sampled individuals are equally distant, corresponding to an ultrametric star tree (meaning each individual is connected to the root of the tree with a single branch; these branches all have the same length).

Looking at this from the opposite angle, if we have samples with a shared evolutionary history but ignore it, we essentially use the wrong ultrametric star tree instead of the true evolutionary tree.

In summary, we assumed full independence of samples in Chapter 4 and, in particular, no linkage between sites (corresponding to an ultrametric star tree). In this chapter, we assumed that samples are interdependent through shared ancestry in the form of a phylogenetic tree, and all the sites are linked. In Chapter 11, we will consider the intermediate case, where samples have some linked and some unlinked sites, which we will model using networks.

9 Phylodynamics

In the previous chapters, we explored the field of phylogenetics. We assumed that for a set of samples, a phylogenetic tree describes their evolutionary relationships, discussed methods to infer this phylogenetic tree, and studied genotypic and phenotypic evolution along the tree. Importantly, these evolutionary processes were assumed not to affect tree generation; instead, they happen along a given tree. A key goal in phylogenetics is to infer the past "state" of the population of interest, the phylogenetic tree, based on the genetic sequences and the evolutionary models.

The size and genetic composition of a population through time are of core interest. Classically, population size and its changes through time are studied by population dynamics (going back to Lotka (1910) and Volterra (1928)). In contrast, genetic changes within a population of a given size (its evolution) are studied by the field of population genetics (going back to Fisher (1930), Wright (1931) and Haldane (1932); and later to the coalescent framework (Kingman 1982)).

In what follows, we introduce *phylodynamics*, a term coined by Grenfell et al. (2004), to consider both population dynamics and population genetics based on a sample of genetic sequences. In particular, phylodynamics studies how phylogenetic trees are generated and which factors shape the trees.

The tree generation process is a population dynamic process of individuals representing some biological unit, and these individuals are replicating and dying. For example, in macroevolution, trees are generated as species undergo speciation, which induces new branching events, and extinction, which induces branch termination. Analogously, in infectious disease epidemiology, infected individuals induce new branching events through disease transmission, and branches terminate upon recovery or death. For more examples, see Section 1.1.1. The goal of phylodynamics is to infer properties of the past "process" of tree generation (rather than the past "state" as in phylogenetics).

A wide variety of factors can influence tree generation processes. For example, for viruses and other rapidly evolving pathogens, phylogenetic trees are shaped by evolutionary, epidemiological and immunological dynamics, and phylodynamics aims to quantify these processes. To consider this in more detail, we note that viruses typically have large population sizes, high mutation rates, and short generation times, producing evolutionary rates so fast that the virus and its genome may undergo significant changes over the course of a single epidemic. The evolution of the virus happens within infected individuals and at transmission bottlenecks. Evolutionary change can be neutral or can happen in response to host immune systems. This means that virus evolutionary, epidemiological, and immunological processes operate on similar timescales. Consequently, viral genetic sequences — and the corresponding phylogenetic tree we use to interpret them — may contain evidence of the epidemiological and immunological dynamics that influenced the evolution of the genetic diversity they represent.

Throughout this chapter, we will assume that we have already obtained the phylogenetic tree from the data, with branch lengths representing calendar time, for example, days, months, or years (see Section 6.4.3). Such trees are called time trees. We will discuss the two main modelling frameworks for generating such trees, namely birth-death and coalescent models, and how vital parameters such as speciation and transmission rates or changes in population sizes can be estimated from the phylogenetic tree under these frameworks. Estimating phylodynamic parameters simultaneously with the phylogenetic tree will be the topic of Chapter 10.

9.1 Birth-death models

An important class of models used in phylodynamic analyses are the *birth-death models*, in which birth and death events give rise to the phylogenetic tree. Phylodynamic analysis using these birth-death models aims to understand and quantify the birth and death rates in the studied population based on a phylogenetic tree. Mathematical derivations throughout this section — which sometimes involve lengthy algebraic expressions — are illustrated in the accompanying Mathematica file.

9.1.1 Population dynamic model

The basic population dynamic model, the *constant rate birth-death model*, is shown in Figure 9.1. In this model, the compartment labelled I, which stands for "Individuals," represents the population. Individuals may correspond to any biological units discussed in Section 1.1.1. In what follows, we assume that the process starts at some time 0 with $I(0)$ individuals, which is the initial state. All individuals in this population are identical and give rise to other individuals at a birth rate β and die at a death rate δ. We call the compartments within a model the "states", and the rates quantify the "dynamics". Such models are called *compartmental models*.

Throughout this section, we will explain phylodynamic principles based on this basic model and mention how time dependence or competition between individuals is introduced into the models at the end of the section. Many applications require models with sub-populations to take into account heterogeneity across individuals. We will discuss such models in Section 9.5.

A compartmental model can be considered in a deterministic or stochastic manner. In the deterministic formulation, where $I(t)$ denotes the number of individuals in compartment I

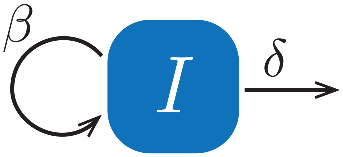


Figure 9.1: Constant rate birth-death model. Individuals from compartment I are born at birth rate β and die at death rate δ.

after time t has elapsed, the change in $I(t)$ is described as

$$\frac{\mathrm{d}}{\mathrm{d}t}I(t) = (\beta - \delta)I(t). \tag{9.1}$$

Thus, since $\frac{\mathrm{d}}{\mathrm{d}t}e^{(\beta-\delta)t} = (\beta - \delta)e^{(\beta-\delta)t}$,

$$I(t) = I(0)e^{(\beta-\delta)t} \tag{9.2}$$

is the solution to this differential equation. If the birth rate is equal to the death rate, the derivative above will be equal to 0, and the size of the population will remain constant through time. If the birth rate is greater than the death rate, the population will grow exponentially, while if the death rate is greater than the birth rate, the population size will tend to zero. Note that $I(t)$ not only takes integer values but can also take any non-integer value between 0 and ∞. This means that $I(t)$ under the deterministic model is not the actual population size but rather the average population size at the given point in time (see also Theorem 9.1.4).

In phylodynamic applications, we do not track population averages as in the deterministic case, but we track individuals represented as branches within the phylogenetic tree. The stochastic formulation of the birth-death model allows us to view the phylogenetic tree as a stochastic outcome of the model.

In the stochastic formulation, the rate r associated with an arrow in the compartmental model means that the probability of the corresponding event happening within a small time step Δt is $r\Delta t$ (see also Section 5.2.2.1). Thus, for the constant rate birth-death model, the probability that an individual gives birth to a new individual in a small time step is $\beta\Delta t$. The probability that an individual dies in a small time step is $\delta\Delta t$. Therefore, the waiting time to a birth event is exponentially distributed with the parameter β, and the waiting time to a death event is exponentially distributed with the parameter δ (see also Section 5.2.2 and Box 17 on page 89). According to Box 21 on page 96, an individual gives birth through time according to a Poisson process with parameter β.

In phylodynamics, we assume that the birth-death model starts with one individual at time 0 and stops after time T, usually representing present time. A simulation of such a process is shown in Figure 9.2 **A**.

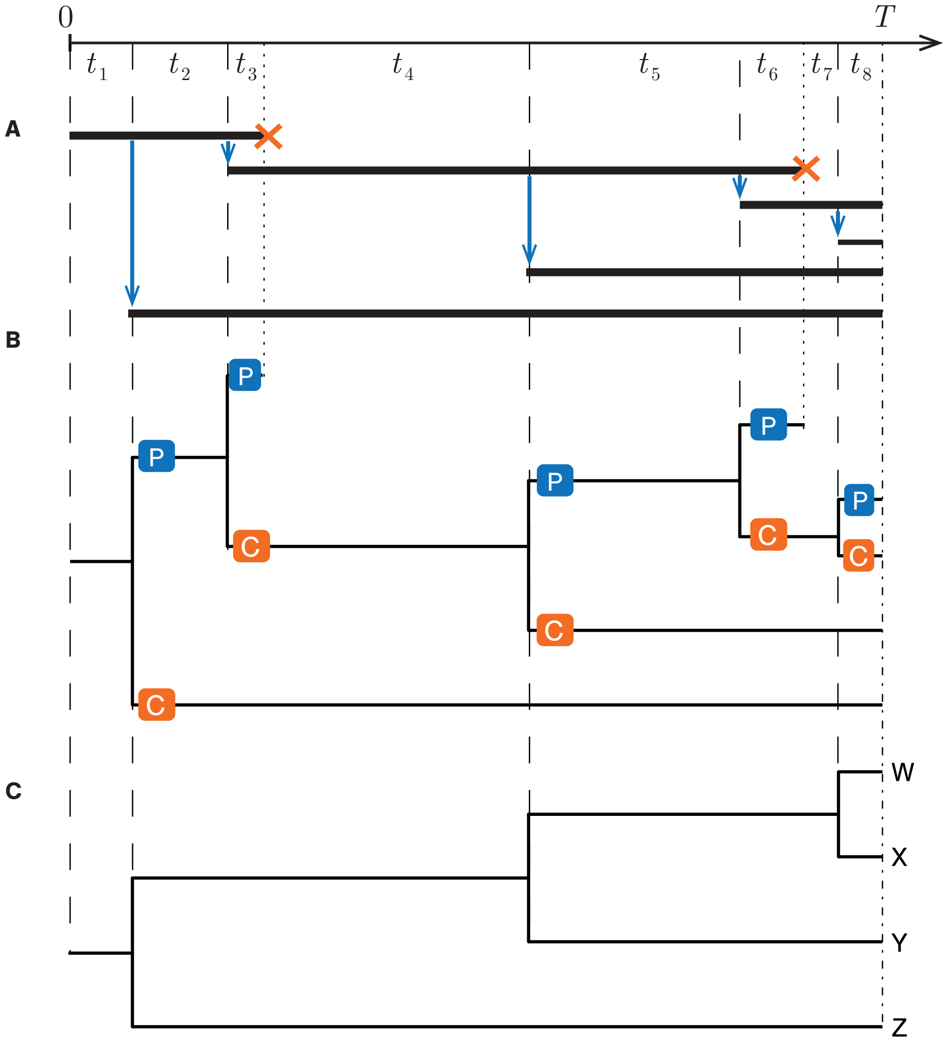


Figure 9.2: **A** Representation of the full population dynamics of one realisation of the birth-death model. Time is shown on the x-axis; each horizontal solid black line in the graph represents an individual's lifetime, blue arrows represent birth events, and orange crosses represent death events. The process is stopped after T timesteps, and the horizontal solid black lines reaching T correspond to individuals alive at time T. Dashed lines indicate birth event times, dotted lines indicate death events times, and the dot-dashed line indicates the end time T of the process. **B** The complete tree of the population dynamics of the top figure (P stands for parent, C for child). **C** Phylogenetic tree of extant lineages resulting from pruning the lineages without descendants at time T from the complete tree.

We will now derive the mathematical properties of this birth-death model, which we will use in a phylodynamic context in Section 9.1.2. In particular, we will derive the population size through time under a birth-death process that starts with a single individual at time 0. Throughout this section, we assume $\beta > \delta > 0$ unless stated otherwise, meaning that on expectation, the process gives rise to an exponentially growing population within which individuals may die. For the cases $\beta = \delta$ and $\delta = 0$, one can take the limit $\delta \to \beta$ and $\delta \to 0$ as done in Stadler and Steel (2019) and the case $0 < \beta < \delta$ is covered in Stadler et al. (2013). We note that many derivations throughout this section also hold for $\beta < \delta$ (see also accompanying Mathematica file); however, for presenting the main concepts, we focus on a single individual giving rise to an exponentially growing population ($\beta > \delta$).

We consider the population size, X_t, after time t has elapsed. The state space of the random variable X_t is $\{0, 1, 2, \ldots\}$. Formally, this offspring population size is a stochastic process $(X_t)_{t\in[0,T]}$. We are interested in the distribution of X_t, the probability of population size n after time t, $P(X_t = n)$. We abbreviate this probability as

$$p(n|t) = P(X_t = n). \tag{9.3}$$

Recall that under the deterministic model, the population size after a time interval t is $e^{(\beta-\delta)t}$. In what follows, we will derive $p(n|t)$ for all n and all t using difference equations leading to differential equations for $p(n|t)$. Solving the differential equations yields a closed-form expression for $p(n|t)$ ($n = \{0, 1, \ldots\}$). These expressions were initially derived using generating functions in Kendall (1948). We employ the expressions when considering the phylogenetic trees generated by birth-death models.

9.1.1.1 Deriving the probability of extinction, $p(0|t)$

The birth-death model starts with one individual. What is the probability that no individuals survived after time interval t, $p(0|t)$?

First, note that for $t = 0$, we have $p(0|t = 0) = 0$ since we assume that the process starts with one individual.

To determine $p(0|t)$ for $t > 0$, we derive its differential equation. In particular, we first derive $p(0|t + \Delta t)$, the probability that the process goes extinct after time $t + \Delta t$, as a function of $p(0|t)$. The process starts with a single individual for which, in a time interval Δt, a death event happens with probability $\delta\Delta t$, and a birth event happens with probability $\beta\Delta t$.

We partition the time interval $t + \Delta t$ as illustrated in Figure 9.3. During the time Δt after the start of the process, four things can happen:

(i) nothing, which has the probability $1 - (\beta + \delta)\Delta t$; then, the original individual has to go extinct within the remaining time t, which has probability $p(0|t)$;

(ii) the individual dies with probability $\delta\Delta t$;

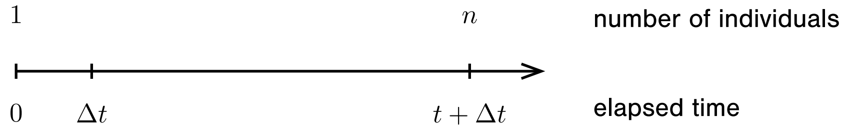


Figure 9.3: Partitioning of time as employed in the derivation of $p(0|t)$ and $p(1|t)$.

(iii) the individual gives birth to another individual with probability $\beta\Delta t$, and both individuals go extinct within time t, which has probability $p(0|t)^2$;

(iv) more than one event happens, which has the probability $\mathcal{O}(\Delta t^2)$.

Thus, we obtain the equation:

$$p(0|t+\Delta t) = \underbrace{(1-(\beta+\delta)\Delta t)p(0|t)}_{(i)} + \underbrace{\delta\Delta t}_{(ii)} + \underbrace{\beta\Delta t p(0|t)^2}_{(iii)} + \underbrace{\mathcal{O}(\Delta t^2)}_{(iv)}. \tag{9.4}$$

Rearranging leads to the following difference equation:

$$\frac{p(0|t+\Delta t)-p(0|t)}{\Delta t} = -(\beta+\delta)p(0|t)+\delta+\beta p(0|t)^2+\mathcal{O}(\Delta t). \tag{9.5}$$

Taking the limit $\Delta t \to 0$ leads to the differential equation:

$$\frac{\mathrm{d}}{\mathrm{d}t}p(0|t) = -(\beta+\delta)p(0|t)+\delta+\beta p(0|t)^2. \tag{9.6}$$

This differential equation is a master equation describing the time evolution of the probability for extinction (see also Section 5.2.2.3).

The solution to this differential equation with the initial condition $p(0|t=0)=0$ is

$$p(0|t) = \frac{\delta(1-e^{-(\beta-\delta)t})}{\beta-\delta e^{-(\beta-\delta)t}}, \tag{9.7}$$

which can easily be verified by differentiating the expression and plugging it into the differential equation. Furthermore, the formula yields 0 for $t=0$, meaning the initial condition is met.

9.1.1.2 Deriving $p(1|t)$

For $t=0$, we have $p(1|t=0)=1$ since we consider one individual at the start of the process. For $t>0$, we again express $p(1|t+\Delta t)$ as a function of $p(1|t)$ and rely on partitioning time as illustrated in Figure 9.3. Note that for $p(1|t)$, compared to the derivation of $p(0|t)$, a death event cannot occur during the first time step Δt, as then the process would die out and would not lead to one individual at present. We obtain the following equation,

$$p(1|t+\Delta t) = (1-(\beta+\delta)\Delta t)p(1|t) + \beta\Delta t \times 2p(1|t)p(0|t) + \mathcal{O}(\Delta t^2). \tag{9.8}$$

The factor of 2 in this equation accounts for the fact that either one of the descendants of the birth event may lead to the surviving individual after time t. Rearrangement of terms and taking the limit $\Delta t \to 0$ leads to the following differential equation:

$$\frac{\mathrm{d}}{\mathrm{d}t}p(1|t) = -(\beta+\delta)p(1|t) + 2\beta p(1|t)p(0|t). \tag{9.9}$$

The solution to this differential equation with initial condition $p(1|t=0)=1$ is

$$p(1|t) = (1-p(0|t))(1-\frac{\beta}{\delta}p(0|t)), \tag{9.10}$$

which again can easily be verified by (i) differentiating the expression and plugging it into the differential equation, and (ii) evaluating the formula for $t=0$.

9.1.1.3 Deriving $p(n|t)$

Theorem 9.1.1. *The probability for an individual to produce* $n \in \{0,1,2,\ldots\}$ *extant individuals after time* t*,* $p(n|t)$*, is*

$$p(0|t) = \frac{\delta(1-e^{-(\beta-\delta)t})}{\beta-\delta e^{-(\beta-\delta)t}}, \tag{9.11}$$

$$p(1|t) = (1-p(0|t))(1-\frac{\beta}{\delta}p(0|t)), \tag{9.12}$$

$$p(n|t) = p(1|t)\left(\frac{\beta}{\delta}p(0|t)\right)^{n-1} \qquad \textit{for } n \geq 2. \tag{9.13}$$

Proof. The expressions for $n=0$ and $n=1$ have been derived above. To prove the expression for $p(n|t)$, $n \geq 2$, we first note that at time $t=0$ we have one individual, and thus $p(n|t=0)=0$ for $n \geq 2$. Indeed, Equation (9.13) is 0 for $t=0$ and $n>1$, since $p(0|t=0)=0$.

For $t>0$, we again derive the differential equation. We now consider $p(n|t+\Delta t)$ as a function of $p(n|t)$ for all $n \geq 1$. In contrast to the derivation of $p(0|t)$ and $p(1|t)$, we split up time into an interval of length t, followed by a time interval of length Δt as illustrated in Figure 9.4.

To arrive at n ($n \geq 1$) individuals after time $t+\Delta t$, after time t we may arrive at:

(i) n individuals (probability $p(n|t)$) followed by no event in the last time interval (probability $1-n(\beta+\delta)\Delta t$);

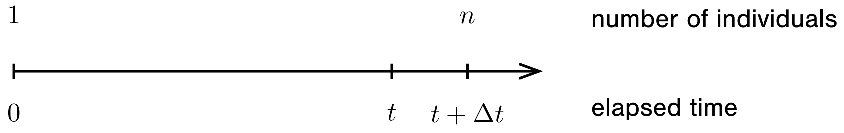


Figure 9.4: Partitioning of time as employed in the derivation of $p(n|t)$.

(ii) $n-1$ individuals (probability $p(n-1|t)$) followed by a birth event in the last interval (probability $(n-1)\beta\Delta t$);

(iii) $n+1$ individuals (probability $p(n+1|t)$) followed by a death event in the last interval (probability $(n+1)\delta\Delta t$);

(iv) any number >0 of individuals, followed by more than two events in the last interval (probability on the order of $\mathcal{O}(\Delta t)^2$).

This leads to the differential equation for $p(n|t)$:

$$\frac{\mathrm{d}}{\mathrm{d}t}p(n|t) = -n(\beta+\delta)p(n|t) + (n-1)\beta p(n-1|t) + (n+1)\delta p(n+1|t) \text{ for } n \geq 1. \quad (9.14)$$

We can now prove the expression for $p(n|t)$ by induction:

Hypothesis to prove: $p(1|t)\left(\frac{\beta}{\delta}p(0|t)\right)^{n-1}$ for $n \geq 2$ is a solution to the differential Equation (9.14).

Base step: Check that the hypothesis holds for $n=2$.
Consider Equation (9.14) for $n=1$: $\frac{\mathrm{d}}{\mathrm{d}t}p(1|t) = -(\beta+\delta)p(1|t) + 2\delta p(2|t)$. Rearranging leads to

$$\begin{aligned} p(2|t) &= \frac{1}{2\delta}\left(\frac{\mathrm{d}}{\mathrm{d}t}p(1|t) + (\beta+\delta)p(1|t)\right) \\ &\overset{(9.9)}{=} \frac{1}{2\delta}\Big(-(\beta+\delta)p(1|t) + 2\beta p(1|t)p(0|t) + (\beta+\delta)p(1|t)\Big) \\ &= p(1|t)\left(\frac{\beta}{\delta}p(0|t)\right). \end{aligned} \quad (9.15)$$

Induction hypothesis: Suppose the hypothesis holds for all $k \leq n$.

Inductive step: Show that the formula holds for $k=n+1$. We consider Equation (9.14) in a rearranged form:

$$(n+1)\delta p(n+1|t) = \frac{\mathrm{d}}{\mathrm{d}t}p(n|t) + n(\beta+\delta)p(n|t) - (n-1)\beta p(n-1|t). \quad (9.16)$$

Differentiating the expression for $p(n|t)$ as stated in Equation (9.13) and combining the result with the expressions in Equations (9.6) and (9.9), we obtain

$$
\begin{aligned}
(n+1)\delta p(n+1|t) &\overset{(9.13)}{=} \frac{\mathrm{d}}{\mathrm{d}t}p(1|t)\left(\frac{\beta}{\delta}p(0|t)\right)^{n-1} + p(1|t)(n-1)\left(\frac{\beta}{\delta}p(0|t)\right)^{n-2}\frac{\beta}{\delta}\frac{\mathrm{d}}{\mathrm{d}t}p(0|t) \\
&\quad + n(\beta+\delta)p(1|t)\left(\frac{\beta}{\delta}p(0|t)\right)^{n-1} - (n-1)\beta p(1|t)\left(\frac{\beta}{\delta}p(0|t)\right)^{n-2} \\
&\overset{(9.6, 9.9)}{=} (-(\beta+\delta)p(1|t) + 2\beta p(1|t)p(0|t))\left(\frac{\beta}{\delta}p(0|t)\right)^{n-1} \\
&\quad + p(1|t)(n-1)\left(\frac{\beta}{\delta}p(0|t)\right)^{n-2}\frac{\beta}{\delta}(-(\beta+\delta)p(0|t) + \delta + \beta p(0|t)^2) \\
&\quad + n(\beta+\delta)p(1|t)\left(\frac{\beta}{\delta}p(0|t)\right)^{n-1} - (n-1)\beta p(1|t)\left(\frac{\beta}{\delta}p(0|t)\right)^{n-2} \\
&= \left(\frac{\beta}{\delta}p(0|t)\right)^{n-2}((n-1)\beta p(1|t) - (n-1)\beta p(1|t)) \\
&\quad + \left(\frac{\beta}{\delta}p(0|t)\right)^{n-1}(-(\beta+\delta)p(1|t) + (-(\beta+\delta)p(1|t)(n-1)) + n(\beta+\delta)p(1|t)) \\
&\quad + \left(\frac{\beta}{\delta}p(0|t)\right)^{n}(2\delta p(1|t) + \delta p(1|t)(n-1)) \\
&= \left(\frac{\beta}{\delta}p(0|t)\right)^{n}\delta(n+1)p(1|t).
\end{aligned}
\tag{9.17}
$$

Thus, we obtain,

$$
p(n+1|t) = p(1|t)\left(\frac{\beta}{\delta}p(0|t)\right)^{n}, \tag{9.18}
$$

which establishes the induction step.

□

It directly follows from Theorem 9.1.1 that,

Corollary 9.1.2. *Consider one individual at some time point. The number of extant descendants produced by this individual after time t, conditioned on non-extinction of the process, has probability function* $\frac{p(n|t)}{1-p(0|t)} = (1-\frac{\beta}{\delta}p(0|t))\left(\frac{\beta}{\delta}p(0|t)\right)^{n-1}$. *This is a geometric distribution with parameter* $\left(1-\frac{\beta}{\delta}p(0|t)\right)$ *(see Box 15 on page 86). Thus,* $\frac{1}{\left(1-\frac{\beta}{\delta}p(0|t)\right)}$ *is the expected number of lineages arising from a single lineage within time t, conditioned that the process survives.*

9.1.2 Phylodynamic model

The simulation displayed in Figure 9.2 **A** does not look like a phylogenetic tree. We obtain the *complete tree* by plotting a branching event for each birth event instead of the blue arrows, as shown in Figure 9.2 **B**. The parent-child relationships are depicted as labels on each branching event: the lineage that already existed before the corresponding birth event is labelled (P) for the parent, and the lineage that just appeared is labelled (C) for the child.

This complete tree contains all lineages that have ever existed during the process. Together with the P/C labels, it represents the total population dynamic history. In most empirical datasets, however, we do not sample the entire population or know the P/C labels.

We now add a *sampling model* to the population dynamics model. Both models together define the *phylodynamic model* giving rise to time trees. First, we define *extant sampling*, which models the sampling of individuals at present. Under the simplest model, each individual at present is sampled with probability ρ.

Second, we define *sampling through time*. The rate of sampling an individual prior to the present time is denoted ψ. Upon sampling, an individual dies with probability r and continues to live with probability $1-r$. Thus, ψr is the death rate with sampling compared to δ, the death rate without sampling. The samples obtained through time may stem from fossils, ancient DNA, or patients throughout an epidemic.

Thus, our full phylodynamic model contains the following parameters (Stadler et al. 2011):

(i) β: birth rate,

(ii) δ: death rate,

(iii) T: time after which the process is stopped,

(iv) ρ: extant sampling probability,

(v) ψ: sampling rate of individuals before the present,

(vi) r: death probability upon sampling of individuals sampled before the present.

Under this model, we can simulate complete trees with sampling events forwards in time. To obtain the phylogenetic tree from a complete tree, we remove the parent-child labels and all lineages without sampled descendants. Then, each sample in the tree is assigned a unique label to obtain the phylogenetic tree.

For $\rho = 1,\ \psi = 0$, a resulting phylogenetic tree is shown in Figure 9.2 **C**, with the samples (the four extant individuals) labeled W, X, Y, Z. For $\rho = 0,\ \psi > 0,\ r = 1$, we observe the phylogenetic tree shown in Figure 9.5 **A**, with the samples (the five individuals sampled through time) labelled A, B, C, D, E. We note that we have a branch ancestral to the root in these phylogenetic trees, which has not been the case for most trees encountered thus far. The start of the branch above the root is the start of the birth-death process (this may correspond to the start of an epidemic or the stem age of a group of species). See also Section 6.2 for the definitions of rooted trees.

For $\rho = 1,\ \psi > 0,\ r < 1$, we observe the phylogenetic tree shown in Figure 9.5 **B**. This latter tree has so-called *sampled ancestors* (Gavryushkina et al. 2014) (labelled with E and F), meaning samples which give rise to further samples. Sampled ancestors can be generated when samples are not removed from the population upon sampling through time (which

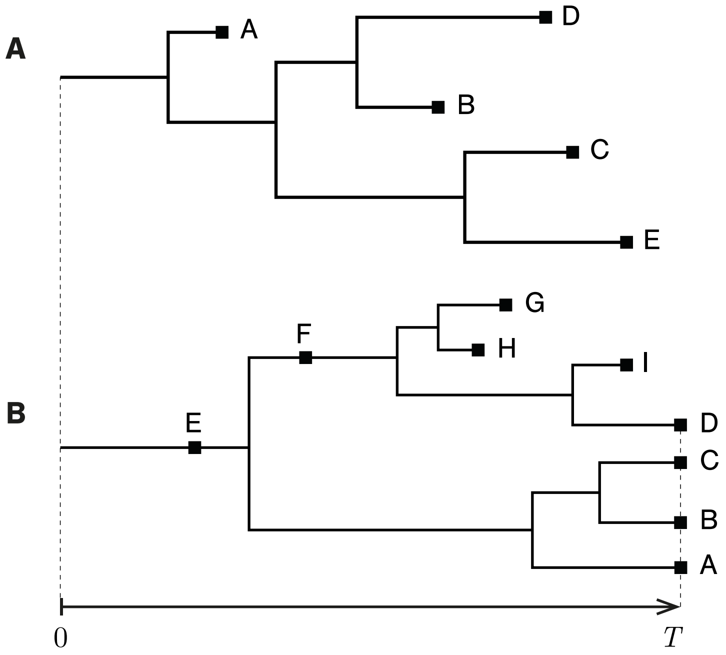


Figure 9.5: Trees generated under the birth-death phylodynamic model. **A** Tree generated with $\rho = 0$, $\psi > 0$ and $r = 1$. We obtain tip samples through time labeled by $A - E$, but no present-day samples. **B** Tree generated with $\rho = 1$, $\psi > 0$ and $r < 1$. We obtain both present-day samples (with labels $A - D$), sampled ancestor samples (with labels $E - F$), and tip samples before the present (with labels $G - I$).

occurs with probability $1 - r$). If a descendant of such a non-removed sample is also sampled, we obtain a sampled ancestor. Note that a labelled tree is now defined as a tree where a unique label is assigned to each sample (tip or sampled ancestor). A sampled ancestor in the context of epidemiology is observed if a patient — after being sampled — infects another patient who is also sampled. Furthermore, fossils within species trees may be sampled ancestors, as descendants of the species representing a fossil sample may be sampled. Note that we did not account for sampled ancestors in Chapter 6 on phylogenetic inference. Instead, all samples were tips in the tree. However, recently published Bayesian inference tools (see Chapter 10) allow for inference with sampled ancestors (Gavryushkina et al. 2017; Zhang et al. 2016).

In what follows, we explain how to infer the population's birth and death rates based on the phylogenetic tree reconstructed from sampled individuals, assuming the introduced phylodynamic birth-death model. Throughout the remainder of the section, we will assume that we sample the entire population at present, that is, $\rho = 1$, but we have no samples from the past, that is, $\psi = 0$ (and thus r is irrelevant). Finally, we continue to assume $\beta > \delta > 0$. This setting allows us to get an intuitive understanding of the phylodynamic models and inference. Finally, we briefly explain how to extend the discussed inference framework to more general scenarios.

9.1.3 Ranked labelled tree topologies

We start by considering a phylogenetic tree where we ignore branch lengths, meaning we only consider the discrete part of the phylogenetic tree, that is, the tree topology. We will show in this section that the topology contains no information about birth and death rates. Later in this chapter, this result will allow us to estimate birth and death rates from the branching times, discarding the topology.

In particular, we consider the ranked labelled tree topology (Ford, Matsen and Stadler 2009) generated by a constant rate birth-death model with $\rho = 1$ and $\psi = 0$. The phylogenetic tree is ranked by assigning a rank to each internal node. An internal node obtains the rank i if this node is the ith branching event in the tree. In particular, the root has rank 1, and the highest rank in a tree on n tips is $n-1$ (see Figure 9.6). A ranked labelled tree topology (discrete part) together with the vector of branching times $x_1, x_2, \dots$ (time between start of process and branching) associated with the nodes of rank $1, 2, \dots$ (continuous part, Figure 9.6) uniquely determine the phylogenetic tree.

Under the constant rate birth-death model, if an event (birth or death) happens, each individual has the same probability of undergoing this event because all individuals have the same birth and death rates. This observation leads to the following theorem.

Theorem 9.1.3. *The constant rate birth-death model with $\rho = 1$ and $\psi = 0$ induces a uniform distribution on ranked labelled tree topologies. Each ranked labelled tree topology has the probability $\frac{2^{n-1}}{n!(n-1)!}$.*

Proof. Consider a realisation of a birth-death model leading to n extant tips where an internal node with rank i has the branching time x_i. Consider the resulting ranked labelled tree topology. To calculate the probability of that ranked labelled tree topology within the set of all such topologies with the same associated branching times $x_1, \dots, x_{n-1}$, we trace the individuals from the present back in time. At time x_{n-1}, we have $\binom{n}{2}$ possibilities to coalesce (merge) two individuals. Since all individuals follow the same dynamics (that is, have the same birth and death rates), the probability of observing the branching event in our given tree is $1/\binom{n}{2}$. We proceed with the same reasoning until we reach the root of the tree. Overall, the probability of the ranked labelled tree topology given any $x_1, \dots, x_{n-1}$ is $\prod_{i=2}^{n} 1/\binom{i}{2}$. Using the definition of the binomial coefficient, we can simplify this expression to

$$\prod_{i=2}^{n} \frac{1}{\binom{i}{2}} = \prod_{i=2}^{n} \frac{2!(i-2)!}{i!} = \frac{2^{n-1}}{n!(n-1)!}. \tag{9.19}$$

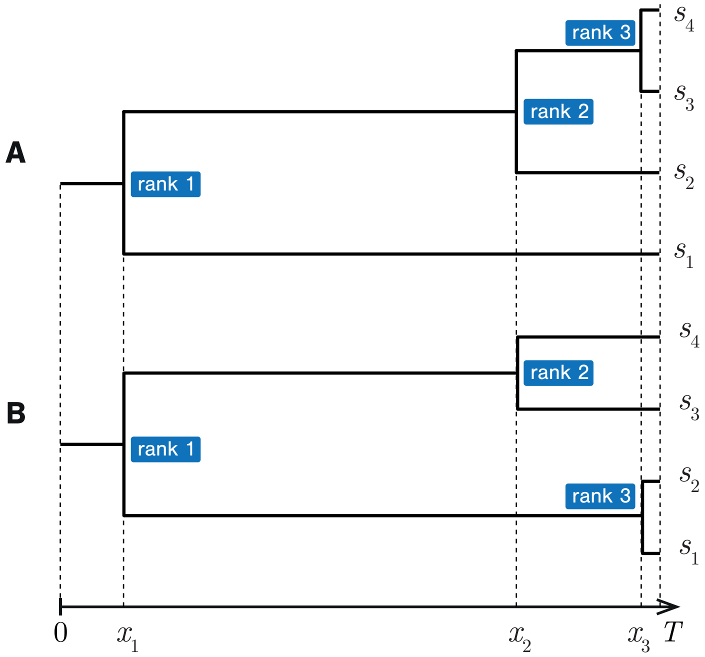


Figure 9.6: A, B Two examples of ranked labelled trees with four tips. The branching times x_1, x_2, x_3 are the same in both trees, but the ranked labelled tree topologies are not.

Thus, the birth-death model induces a uniform distribution on ranked labelled tree topologies on n tips. □

This result implies that the distribution of ranked labelled tree topologies is independent of the birth and death rates. Aldous (2001) generalises this result using reasoning analogous to the proof above. Any model where the birth and death rates are the same across all individuals at every point in time induces a uniform distribution on ranked labelled tree topologies — even if, for example, these rates change through time (Stadler 2011; Morlon, Parsons and Plotkin 2011) or are a function of the overall number of individuals (Etienne et al. 2011). Such models are also called *homogeneous models* as all individuals at the same time point undergo the same dynamics (while this dynamic may differ for different time points). The individuals are called *exchangeable* as they are all equivalent with respect to the population dynamic process. We emphasise that homogeneity here refers to homogeneous individuals at one time point. In Box 24 on page 98 describing Markov chains, we encountered a different type of homogeneity, namely homogenous probabilities through time.

Consequently, when we aim to quantify birth and death rates under models where these rates are the same across co-existing individuals, the ranked labelled tree topology contains no information about these rates. Thus, the branching times $x_1, \ldots, x_{n-1}$ contain all the information about the birth and death rates. For the introduced phylodynamic birth-death model, we will now first provide an intuitive approach for birth-death parameter estimation based on expectations derived from branching times and then a maximum likelihood approach based on the full distribution of branching times.

9.1.4 Expected population sizes and branching times

The plot with the number of lineages through time in a tree on the y-axis versus time on the x-axis is called the *lineages-through-time (LTT) plot*. An example of an LTT plot is shown in Figure 9.7 C. The LTT plot of the complete tree (referred to as the complete LTT plot) shows the population size through time. In contrast, the LTT plot of the phylogenetic tree (referred to as the phylogenetic LTT plot) shows the number of lineages surviving to the present through time. Note that the LTT plot only summarises the branching times but does not display any information regarding the ranked labelled tree topology.

The LTT plots provide one method of estimating the parameters of the birth-death model. To illustrate how this works, we simulate a large number of realisations of the same phylodynamic constant rate birth-death model, obtaining a large number of trees starting with one individual at time 0 and stopping at a fixed time T, using our assumption $\beta > \delta > 0$ and $\rho = 1, \psi = 0$. We then plot the average phylogenetic LTT plot in Figure 9.8, blue; the average complete LTT plot is shown in black. Note that when averaging the LTT plots, all simulations that went extinct prior to time T are ignored as they do not induce a phylogenetic tree. Finally, the average total population size over all simulations, — that is, over the complete trees and the trees that went extinct prior to time T, — is shown with a black dashed line. Note that the y-axis is on the log scale. In what follows, we discuss the shape of these LTT plots under the phylodynamic constant rate birth-death model. The resulting insights show that the LTT plots, the branching times of phylogenetic trees, encode information about both the birth and death rates. In Section 9.1.5, we introduce a likelihood framework for estimating birth and death rates from the branching times in a phylogenetic tree.

9.1.4.1 Expected total population size

In Figure 9.8, we visualise — based on our simulations — that the expected total population size through time has a constant slope $\beta - \delta$ on the log scale (dashed line), corresponding to the expected total population size $e^{(\beta-\delta)t}$ at time t. We denote the expected total population size at time t by $N(t)$ and note that $N(t)$ is the expectation of the random variable X_t (the offspring population size) defined in Section 9.1.1, $N(t) = \mathrm{E}(X_t)$. We prove in the following theorem that $N(t)$ indeed grows exponentially at rate $\beta - \delta$.

Theorem 9.1.4. *The expected number of lineages at time t, $N(t)$, is, for $\beta > \delta > 0$,*

$$N(t) = e^{(\beta-\delta)t}. \tag{9.20}$$

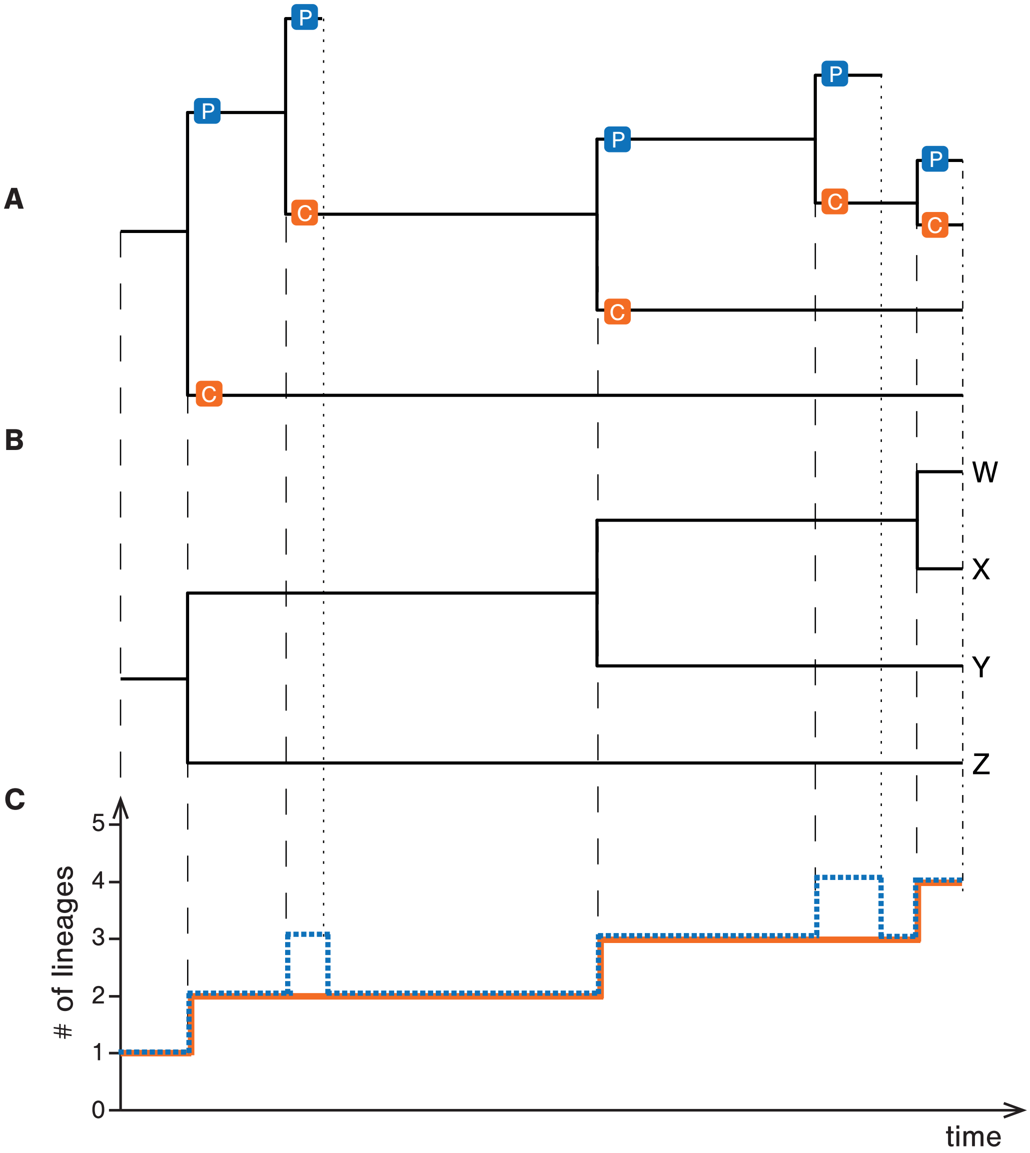


Figure 9.7: **A** Complete tree, **B** phylogenetic tree, and **C** corresponding lineages-through-time (LTT) plot, continuing the example in Figure 9.2. The LTT plot of the complete tree is shown in dotted blue, and the LTT plot of the phylogenetic tree is shown in red.

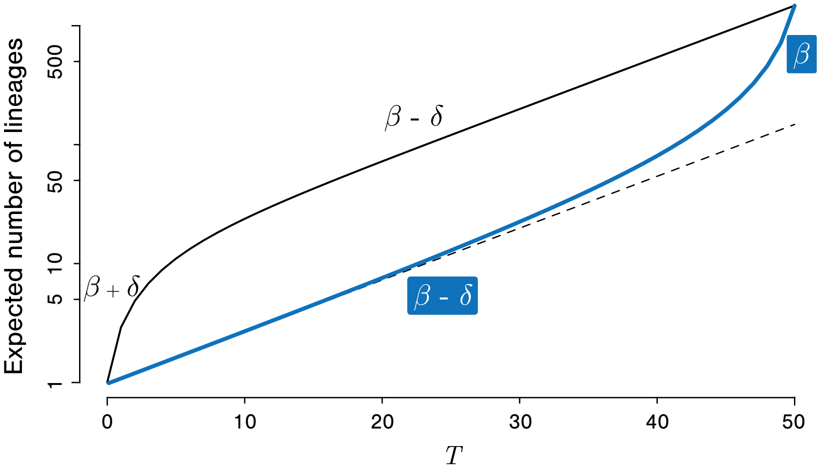


Figure 9.8: Average complete lineages-through-time (LTT) plot (top solid black line) and average phylogenetic LTT plot (bottom blue line) obtained from simulations. Note that in the LTT average, all simulations in which no individual survived to time $T = 50$ are ignored. The dashed black line shows the average total population size (including all simulations that went extinct prior to T). The plots display the scenario $\beta > \delta > 0$ (with $\beta = 1, \delta = 0.9$), that is, an on-average increasing population size where death may occur. The increased slope at the start of the average complete LTT plot is called the *push-of-the-past*; the increased slope at the end of the average phylogenetic LTT plot is called the *pull-of-the-present* (Nee et al. 1994). Note that the slopes $\beta - \delta$ and $\beta + \delta$ are exact quantifications only in the limit $T \to \infty$ while the slope β is exact for all T.

Proof. First, we note that $\sum_{n=1}^{\infty} nx^{n-1} = \frac{1}{(1-x)^2}$ for $-1 < x < 1$. This follows directly from differentiating the formula for an infinite geometric series $\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$, where $-1 < x < 1$. Second, we note that $0 \leq \frac{\beta}{\delta} p(0|t)$, since β, δ and $p(0|t)$ are non-negative. Third, we have

$\frac{\beta}{\delta}p(0|t) = \frac{\beta-\beta e^{-(\beta-\delta)t}}{\beta-\delta e^{-(\beta-\delta)t}} < 1$. Combining these ideas we get $0 \leq \frac{\beta}{\delta}p(0|t) < 1$, and

$$\begin{aligned} N(t) &= \sum_{n=0}^{\infty} np(n|t) \\ &\overset{(9.13)}{=} \sum_{n=1}^{\infty} np(1|t)\left(\frac{\beta}{\delta}p(0|t)\right)^{n-1} \\ &\overset{(9.12)}{=} \frac{p(1|t)}{\left(1-\frac{\beta}{\delta}p(0|t)\right)^2} = \frac{1-p(0|t)}{1-\frac{\beta}{\delta}p(0|t)} \\ &\overset{(9.11)}{=} e^{(\beta-\delta)t}. \end{aligned} \tag{9.21}$$

The last equation follows when using the equality $1-\frac{\beta}{\delta}p(0|t) = e^{-(\beta-\delta)t}(1-p(0|t))$, which can be established from Equation (9.11). □

Note that the expected population size $N(t)$ equals the population size under the deterministic model, $I(t)$.

9.1.4.2 Expected complete LTT plot

Based on the simulation result in Figure 9.8, the complete LTT plot (black) goes through a period of accelerated growth at the beginning of the process before growing at a constant rate. Recall that the complete LTT plot only includes populations that survive to the present. An intuitive explanation for the initial fast growth is that populations that grow slowly at the start are more likely to go extinct before the end of the process and, thus, are not included in the average complete LTT plot. This phenomenon is called the *push-of-the-past* (Nee et al. 1994). In the following, we provide the full equation for the black line, $N_T(t)$, which was originally stated in Harvey, May and Nee (1994) and Nee, May and Harvey (1994).

Theorem 9.1.5. *The expected number of lineages at time $t \leq T$, conditioned on non-extinction at present time T, is denoted $N_T(t) = \mathrm{E}(X_t|X_T > 0)$. We have*

$$N_T(t) = \frac{e^{(\beta-\delta)t}}{1-p(0|T)} - \frac{p(0|T)-p(0|t)}{(1-p(0|t))(1-p(0|T-t))}. \tag{9.22}$$

Proof. We have

$$\begin{aligned}P(X_t = n|X_T > 0) &= \frac{P(X_t = n, X_T > 0)}{P(X_T > 0)}\\ &= \frac{P(X_T > 0|X_t = n)P(X_t = n)}{P(X_T > 0)}\\ &= \frac{(1 - p(0|T - t)^n)p(n|t)}{1 - p(0|T)}. \end{aligned} \tag{9.23}$$

Taking the expectation, we obtain

$$\begin{aligned}N_T(t) &= \sum_{n=1}^{\infty} nP(X_t = n|X_T > 0)\\ &= \frac{\sum_{n=1}^{\infty} np(n|t)}{1 - p(0|T)} - \frac{\sum_{n=1}^{\infty} np(0|T - t)^n p(n|t)}{1 - p(0|T)}. \end{aligned} \tag{9.24}$$

Using Theorem 9.1.4 for the left expression and $\sum_{n=1}^{\infty} nx^{n-1} = \frac{1}{(1-x)^2}$ for the right expression (see also the proof of Theorem 9.1.4 for the latter), we obtain

$$N_T(t) = \frac{e^{(\beta-\delta)t}}{1 - p(0|T)} - \frac{p(1|t)p(0|T - t)}{1 - p(0|T)} \frac{1}{(1 - \frac{\beta}{\delta}p(0|t)p(0|T - t))^2}. \tag{9.25}$$

This can be simplified to Equation (9.22) as also demonstrated in the accompanying Mathematica file.

□

To investigate the shape of the LTT plot under the constant rate birth-death model, we consider the derivative of $\log(N_T(t))$. The derivative is the slope of the LTT plot. We obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\log(N_T(t)) = \frac{\beta(\beta-\delta)\left(\delta + \beta e^{2t(\beta-\delta)} - 2\delta e^{(\beta-\delta)(2t-T)}\right)}{-\beta\delta + \beta^2 e^{2t(\beta-\delta)} - 2\beta\delta e^{(\beta-\delta)(2t-T)} + \delta(\beta+\delta)e^{(\beta-\delta)(t-T)}}, \tag{9.26}$$

which simplifies for the cases $t = 0$ and $t = T$ to

$$\frac{\mathrm{d}}{\mathrm{d}t}\log(N_T(t = 0)) = \beta + \delta + \frac{\delta(\beta-\delta)}{\delta - \beta e^{(\beta-\delta)T}}, \tag{9.27}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\log(N_T(t = T)) = \beta - \delta + \frac{\delta(\beta-\delta)^2}{(\delta - \beta e^{(\beta-\delta)T})^2}. \tag{9.28}$$

We now discuss the shape of the expected complete LTT plot in the limit; recall that we assume

$\beta > \delta > 0$. First, we note that

$$\lim_{T \to 0} \frac{\mathrm{d}}{\mathrm{d}t} \log(N_T(t=0)) = \beta, \tag{9.29}$$

and is monotonously increasing with T, leading to

$$\lim_{T \to \infty} \frac{\mathrm{d}}{\mathrm{d}t} \log(N_T(t=0)) = \beta + \delta. \tag{9.30}$$

Thus, the initial slope of the complete LTT plot is between β and $\beta + \delta$. Furthermore,

$$\lim_{T \to 0} \frac{\mathrm{d}}{\mathrm{d}t} \log(N_T(t=T)) = \beta, \tag{9.31}$$

and is monotonously decreasing with increasing T, leading to

$$\lim_{T \to \infty} \frac{\mathrm{d}}{\mathrm{d}t} \log(N_T(t=T)) = \beta - \delta. \tag{9.32}$$

Thus, the final slope of the complete LTT plot is between β and $\beta - \delta$.

In summary, for $T \to \infty$, the slope of the LTT plot at the start of the process is $\beta + \delta$ and decreases with time to $\beta - \delta$. For finite T, lower initial and higher final slopes are observed, quantified in Equation (9.28). For $T \to 0$, the slope is β.

9.1.4.3 Expected phylogenetic LTT plot

Based on the simulation results in Figure 9.8, blue, the phylogenetic LTT plot initially grows with a constant rate and has an accelerated growth close to the present. This phenomenon is called the *pull-of-the-present* (Nee et al. 1994). An intuitive explanation for the accelerated recent growth is that lineages appearing close to the present have less time to go extinct and, thus, are more likely to be sampled, leading to an apparent increase in the number of lineages in the phylogenetic tree. Again, we provide the equation for the blue line, $N_{T,p}(t)$.

Theorem 9.1.6. *The expected number of lineages through time in a phylogenetic tree conditioned on non-extinction, $N_{T,p}(t)$ is*

$$N_{T,p}(t) = \frac{e^{(\beta-\delta)t}(1 - p(0|T-t))}{1 - p(0|T)}. \tag{9.33}$$

Proof. This result was first presented in Harvey, May and Nee (1994) and Kubo and Iwasa (1995). We prove the result analogously to Kubo and Iwasa (1995).

The expected number of lineages after time t, conditioned on non-extinction, is $1/\left(1-\frac{\beta}{\delta}p(0|t)\right)$ (see Corrolary 9.1.2). Thus,

$$N_{T,p}(T) = \frac{1}{1-\frac{\beta}{\delta}p(0|T)}. \tag{9.34}$$

Furthermore, the average number of offspring at time T produced by one individual at time t, $N_{T,p}(T)/N_{T,p}(t)$, is equivalent to the expected number of offspring of one individual after time $T-t$. Using Corrolary 9.1.2, it follows that

$$\frac{N_{T,p}(T)}{N_{T,p}(t)} = \frac{1}{1-\frac{\beta}{\delta}p(0|T-t)}. \tag{9.35}$$

In summary, we obtain

$$N_{T,p}(t) = \frac{1-\frac{\beta}{\delta}p(0|T-t)}{1-\frac{\beta}{\delta}p(0|T)}. \tag{9.36}$$

Since $1-\frac{\beta}{\delta}p(0|t) = e^{-(\beta-\delta)t}(1-p(0|t))$ (see proof of Theorem 9.1.4), we complete the proof. □

To investigate what the phylogenetic LTT plot looks like under the constant rate birth-death model, we consider the derivative of $\log(N_{T,p}(t))$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\log(N_{T,p}(t)) = \frac{\beta(\beta-\delta)}{\beta-\delta e^{(\beta-\delta)(t-T)}}, \tag{9.37}$$

and thus,

$$\frac{\mathrm{d}}{\mathrm{d}t}\log(N_{T,p}(t=0)) = \frac{\beta(\beta-\delta)}{\beta-\delta e^{-(\beta-\delta)T}}, \tag{9.38}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\log(N_{T,p}(t=T)) = \beta. \tag{9.39}$$

For $t=0$, we further note that $\lim_{T\to 0}\frac{\mathrm{d}}{\mathrm{d}t}\log(N_{T,p}(t=0)) = \beta$, and this function is monotonously decreasing with an increasing T, to reach in the limit $\lim_{T\to\infty}\frac{\mathrm{d}}{\mathrm{d}t}\log(N_{T,p}(t=0)) = \beta-\delta$ (recall that $\beta > \delta > 0$).

In summary, the phylogenetic LTT plot has a slope β at present and decreases going into the past towards $\beta-\delta$. The slope $\beta-\delta$ is reached for $T\to\infty$.

Recall that all our derivations were for constant birth and death rates through time. When birth and death rates are functions of time, the works by Kendall (1948), Nee, May and Harvey (1994) and Kubo and Iwasa (1995) lead to a generalisation of the expressions given in Theorem 9.1.1 and Corrolary 9.1.2 as well as the expressions for the LTT plots, primarily relying on the concept of generating functions.

9.1.4.4 Parameter estimation with LTT plots

Using these insights, we can determine the birth and death rates of an empirical phylogenetic tree by displaying its LTT plot, Figure 9.9: each black square corresponds to a branching event, its time displayed on the x-axis, and the number of lineages in the tree after the branching event is displayed on the y-axis. Given that we observe a phylogenetic tree, we first conclude that the population is growing, thus $\beta > \delta$. Assuming T is large enough, the initial slope of the LTT plot should (on expectation) be $\beta - \delta$, and its recent slope should (on expectation) be β (see Section 9.1.4.3). Thus, in theory, we can estimate the birth and the death rates from the slopes of the two regression lines fitted to the black squares of the empirical LTT plot, as shown in Figure 9.9, bold blue lines. One regression is performed on the early part of the LTT plot to estimate $\beta - \delta$, and one on the late part to estimate β.

However, there are problems with this method of estimating the parameters of the birth-death models. Firstly, the variance in the timing of the next branching event (the next square in Figure 9.9) decreases with increasing population size. Thus, a classic linear regression (see Chapter 8) assuming the same variance for each data point is not valid. Secondly, the time of transition between the two phases of the curve is unclear. This makes it difficult to decide where to place the cutoff between points used to fit the first linear regression line and those used to fit the second linear regression line. Lastly, the value of the initial slope $\beta - \delta$ is derived for $T \to \infty$ and is higher for finite T (Section 9.1.4.3).

Nevertheless, this ad-hoc approach should illustrate that phylogenetic trees, despite not including samples of all individuals (here we exclude all individuals without descendants at time T), provide information on both birth and death rates. The next section explains how to coherently estimate birth and death rates from a phylogenetic tree using a likelihood-based approach.

9.1.5 Distribution of branching times

We now derive the probability density of a labelled phylogenetic tree under the phylodynamic constant rate birth-death model. Given the uniform distribution on ranked labelled trees, this derivation essentially provides the probability density of branching times, whereas, in the previous section, we only considered the expectation of branching times. Based on these derivations, we provide a maximum likelihood approach to estimate the birth and death rates based on the branching times in a phylogenetic tree. Such an approach will overcome the problems mentioned above when using linear regression on LTT plots to estimate birth and death rates.

The birth-death process is stochastic forwards in time. We typically track an individual from time 0 until time T. When considering the tree, we interpret 0 as the present, and no samples are taken later than the present. We then increase time going into the past, assuming the tree has a single individual at some time $T > 0$ in the past.

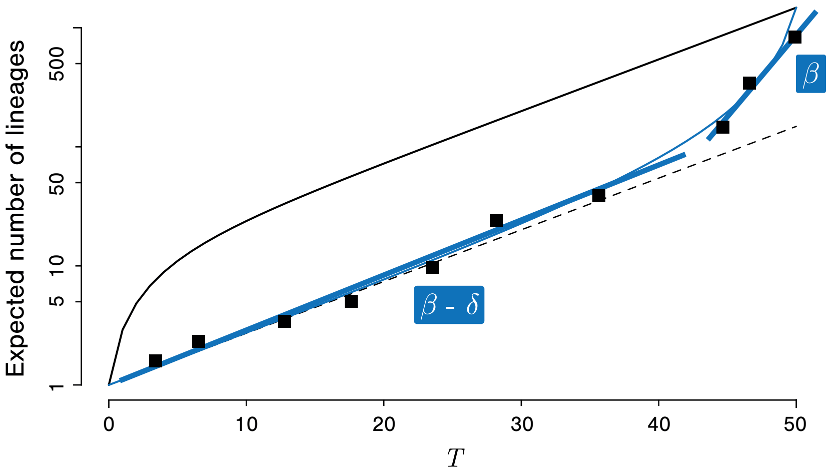


Figure 9.9: The plot shows the fit of two regression lines (blue) to an empirical phylogenetic LTT plot (black squares), on top of the LTT plots from Figure 9.8. We note that the first four squares from the left correspond to the root and the following three branching events. To ensure readability, we do not plot squares for all the later branching events.

9.1.5.1 Parameter estimation with maximum likelihood

Recall that in phylogenetics, we calculate the phylogenetic likelihood $L(\mathcal{T}, Q; D) = P(D|\mathcal{T}, Q)$, for a phylogenetic tree $\mathcal{T}$ and substitution rate matrix Q, given an MSA D. In particular, the tree $\mathcal{T}$ is a parameter. The aim of phylodynamics is to calculate the so-called *phylodynamic likelihood*: $L(\eta = (\beta, \delta, T, \rho, \psi, r); \mathcal{T}) = P(\mathcal{T}|\eta)$ where $\mathcal{T}$ is now the data. Given a fixed phylogenetic tree $\mathcal{T}$, we aim to determine the maximum likelihood estimate for the birth-death parameters (summarised in η). Such a maximum likelihood approach — ignoring sampling through time — was first introduced in Thompson (1975) and then in Nee, May and Harvey (1994). In the following, we derive $P(\mathcal{T}|\eta)$ based on the assumption of complete extant tip sampling, $\rho = 1$ and $\psi = 0$.

Consider a phylogenetic tree $\mathcal{T}$ on n extant tips, which is obtained from a birth-death model stopped at time T. We measure time in reverse for convenience. In particular, we set the present time as 0. The $n - 1$ branching events occur at times $x_1 > x_2 > \ldots > x_{n-1}$ prior to present (note that since our stochastic process is continuous in time with constant rates, the probability of two branching events at the same time is 0). Finally, we define the start time of the process as $x_0 = T$. To facilitate the mathematical derivation, we label the two descendants of each branching event with “left” and “right”, while the tips are not labelled. Such trees are also called *oriented trees*, while the phylogenetic trees considered so far were labelled trees. An example of an oriented tree with four tips is shown in Figure 9.10.

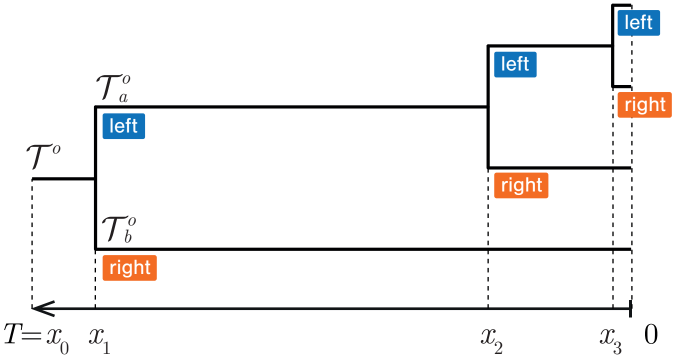


Figure 9.10: Oriented tree $\mathcal{T}^o$ with four tips on which the phylodynamic likelihood calculation is explained. The tree has the two subtrees $\mathcal{T}_a^o$ and $\mathcal{T}_b^o$. We reverse time such that the present time is 0 and the first individual appeared at T time units in the past.

Suppose this example tree evolved under a constant rate birth-death model without death, $\delta = 0$. Then, the probability density of this oriented tree $\mathcal{T}^o$ is a product of exponentials and rates:

$$P(\mathcal{T}^o|\beta, \delta = 0, T = x_0, \rho = 1, \psi = 0, r) = e^{-\beta(x_0 - x_1)}\beta \underbrace{e^{-\beta x_1}}_{\mathcal{T}_b^o} \underbrace{e^{-\beta(x_1 - x_2)}\beta e^{-\beta x_2} e^{-\beta(x_2 - x_3)}\beta e^{-2\beta x_3}}_{\mathcal{T}_a^o}$$

$$= \beta^3 \prod_{i=0}^{3} e^{-\beta x_i}, \qquad (9.40)$$

where $\mathcal{T}_a^o$ is the left and $\mathcal{T}_b^o$ is the right subtree descending branching time x_1 as visualized in Figure 9.10.

For $\beta > \delta > 0$, the probability density calculation is more complicated: we need to take into account all possible unobserved events leading to extinct (meaning unsampled) subtrees. Suppose we were to know all these unobserved events, that is, the complete tree in which the phylogenetic tree is embedded. In that case, we could calculate the probability of the complete tree as a product of exponentials and rates. Since we do not have information on the unobserved events leading to extinct subtrees, we must sum over them. We could sum over the probability density of all complete trees in which our oriented phylogenetic tree is embedded. However, this will be too slow in practice as there are infinitely many such complete trees. Below, we will do this summation using differential equations instead.

The probability density of the example oriented tree in Figure 9.10 can be calculated as the product of the probability density of the first branch between x_0 and x_1, the probability density of a branching event at time x_1, and the probability density of the two subtrees $\mathcal{T}_a^o$ and $\mathcal{T}_b^o$ descending from time x_1. Here, we employ the property of the birth-death model, which is that birth and death events occur independently in different parts of the tree. Let

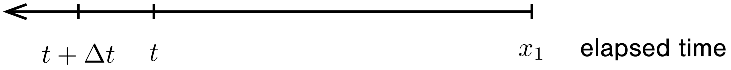


Figure 9.11: Partitioning of time employed in the derivation of $p(t, x_1)$.

$p(x_0, x_1)$ be the probability density of an individual — at time x_0 in the past — to produce a branch of length $x_0 - x_1$. Then, the probability density of an oriented tree $\mathcal{T}^o$ with age $T = x_0$ is:

$$P(\mathcal{T}^o|T = x_0) = p(x_0, x_1)\beta P(\mathcal{T}^o_a|T = x_1)P(\mathcal{T}^o_b|T = x_1). \tag{9.41}$$

Note that we omit the parameters $\beta, \delta, \rho = 1, \psi = 0, r$ in the expression for the probability density of the tree to facilitate readability. We can continue expanding this expression recursively until we come to the tips of the tree, meaning the tree probability density is the product of the probability densities of branches and birth rates.

To calculate the probability density of the branch between t and x_1, $p(t, x_1)$ (where $t \geq x_1$), we write down the equation in discrete time using the same considerations as in the derivation of $p(1|t)$ but reversed in time (as shown in Figure 9.11):

$$p(t + \Delta t, x_1) = (1 - (\beta + \delta)\Delta t)p(t, x_1) + 2\beta\Delta t p(t, x_1)p(0|t) + \mathcal{O}(\Delta t^2), \tag{9.42}$$

which after rearranging and taking the limit $\Delta t \to 0$ gives us the following differential equation:

$$\frac{\mathrm{d}}{\mathrm{d}t}p(t, x_1) = -(\beta + \delta)p(t, x_1) + 2\beta p(t, x_1)p(0|t). \tag{9.43}$$

This is the same differential equation as for $p(1|t)$. The initial condition is $p(x_1, x_1) = 1$, as the presence of an individual in the tree at time x_1 induces with probability one a branch of length zero. Thus, $p(t, x_1) = \frac{p(1|t)}{p(1|x_1)}$, and in particular $p(x_0, x_1) = \frac{p(1|x_0)}{p(1|x_1)}$.

Overall, we obtain,

$$P(\mathcal{T}^o|T = x_0) = p(x_0, x_1)\beta P(\mathcal{T}_a|T = x_1)P(\mathcal{T}_b|T = x_1) = \beta^{n-1}\prod_{i=0}^{n-1} p(1|x_i), \tag{9.44}$$

using the observation that each internal node is once the ending and twice the starting point of a branch. This proves the following theorem:

Theorem 9.1.7. *Consider the constant rate birth-death model for time T with birth rate β and death rate δ. Further, consider complete extant tip sampling ($\rho = 1$) and no sampling through time ($\psi = 0$). The probability density of an oriented tree $\mathcal{T}^o$, conditioned on non-extinction ($X_T > 0$), is*

$$P(\mathcal{T}^o|T = x_0, X_T > 0, \beta, \delta) = \frac{p(1|x_0)}{1 - p(0|x_0)}\prod_{i=1}^{n-1} \beta p(1|x_i). \tag{9.45}$$

Analogous probability densities for different types of conditioning are provided in Stadler (2013).

To obtain the probability density of a labelled phylogenetic tree $\mathcal{T}$ on n samples, we note that we can label the samples of an oriented tree in $n!$ ways, where each labelling has the same probability. Further, we note that given a tree (without sampled ancestors) on $n-1$ internal nodes, for a particular labelling, there are 2^{n-1} orientations (left or right for each of the $n-1$ internal nodes). Thus,

Corollary 9.1.8. *The probability density of a labelled tree $\mathcal{T}$, with $\rho = 1, \psi = 0$ and conditioned on non-extinction, is*

$$P(\mathcal{T}|T = x_0, X_T > 0, \beta, \delta) = \frac{2^{n-1}}{n!} \frac{p(1|x_0)}{1 - p(0|x_0)} \prod_{i=1}^{n-1} \beta p(1|x_i). \tag{9.46}$$

In Theorem 9.1.3, we showed that each ranked labelled tree topology has the same probability. Thus,

Corollary 9.1.9. *The probability density of the branching times $x_1 > x_2 > \ldots > x_{n-1}$, meaning the probability density of the LTT plot, with $\rho = 1, \psi = 0$ and conditioned on non-extinction, is*

$$P(x_1, x_2, \ldots, x_{n-1}|T = x_0, X_T > 0, \beta, \delta) = (n-1)! \frac{p(1|x_0)}{1 - p(0|x_0)} \prod_{i=1}^{n-1} \beta p(1|x_i). \tag{9.47}$$

We can now perform maximum likelihood inference on the parameters β and δ based on the branching times in the phylogenetic tree. The parameter T is typically fixed, as it is the stem age of a group of species in macroevolution or the start of an epidemic in epidemiology. Alternatively, the probability of the branching times conditioned on the first branching event (x_1) can be derived (for an overview, see e.g. Stadler (2013)). The values for β and δ that maximise the probability density $P(x_1, x_2, \ldots, x_{n-1}|T = x_0, X_T > 0, \beta, \delta)$ are the maximum likelihood parameter estimates of the birth-death model for the given branching times. We note that the expressions in Theorem 9.1.7 and Corollaries 9.1.8 and 9.1.9 all lead to the same maximum likelihood parameter estimates, as these expressions only differ in a function that depends on the number of samples, n, but not on the tree. Importantly, this maximum likelihood approach considers the full probability density of the LTT plot, while the linear regression approach from the last section only uses properties of the expected LTT plot.

In this chapter, we quantify the birth and death rates based on a given phylogenetic tree $\mathcal{T}$. In particular, the maximum likelihood birth and death rate estimates are $(\hat{\beta}, \hat{\delta}) = \text{argmax}_{\beta,\delta} P(\mathcal{T}|T = x_0, X_T > 0, \beta, \delta)$. The interested reader may wonder how to quantify the birth and death rates based on sequence data D,

$$(\hat{\beta}, \hat{\delta}) = \text{argmax}_{\beta,\delta} P(D|T = x_0, X_T > 0, \beta, \delta). \tag{9.48}$$

So far, we have not established an expression for $P(D|T = x_0, X_T > 0, \beta, \delta)$. However, we can rewrite:

$$\begin{aligned} P(D|T = x_0, X_T > 0, \beta, \delta) &= \int_{\mathcal{T}} P(D, \mathcal{T}|T = x_0, X_T > 0, \beta, \delta) \, \mathrm{d}\mathcal{T} \\ &= \int_{\mathcal{T}} P(D|\mathcal{T}) P(\mathcal{T}|T = x_0, X_T > 0, \beta, \delta) \, \mathrm{d}\mathcal{T}. \end{aligned} \tag{9.49}$$

In the last equation, we make the common assumption that the sequence evolution process is independent of the tree generation process. While we can calculate the two expressions within the last integral, we cannot analytically integrate over all trees (see Section 6.2.3.3 on the size of the tree space). In Chapter 10, we will discuss numerical algorithms to deal with this integral.

9.1.5.2 General birth-death models

As mentioned above, the probability density of a constant rate birth-death tree with complete present-day sampling ($\rho = 1$) and no sampling through time ($\psi = 0$) was initially calculated in Thompson (1975, page 58). Nee, May and Harvey (1994) further accounted for $\rho < 1$ and time-dependence of the birth and death parameters. In the derivations above, $\rho < 1$ can be accounted for by changing the initial conditions. The structure of the resulting probability density remains the same as in Corrolary 9.1.8; however, the functions $p(0|t)$ and $p(1|t)$ change (Stadler 2013). We can further account for the time dependence by changing the birth and death parameters in the differential equations through time.

Over the past few years, the probability densities for phylogenetic trees have been derived for extensions of this basic birth-death model. The derivations for these generalisations rely on the ideas introduced above.

Allowing for sampling through time ($\psi > 0$) with $r = 0$ (recall that r is defined as the probability for death upon sampling prior to the present) has been introduced in Stadler (2010). Allowing for sampling through time essentially requires changing initial conditions and slightly modifying the differential equations. The birth and death parameters were allowed to change through time with $r = 1$ in Stadler et al. (2013). In particular, the rates changed in a piecewise constant fashion, and this model is also called *birth-death skyline plot*. The extension to time dependence again relies on changing the parameters in the differential equations through time. Furthermore, Stadler et al. (2013) provided the possibility to assume a different sampling probability ρ_i at time points t_i by modifying the initial conditions. This extension can be used, for instance, to represent mass extinctions (in macroevolution datasets) or punctual sampling efforts (in epidemiology). Gavryushkina et al. (2014) used such models with any $r \in [0, 1]$.

For the models thus far, the differential equations describing the tree probability density can be solved analytically. Allowing for competition among co-existing individuals has been accounted for in Leventhal et al. (2014) and Vaughan et al. (2019) requiring the numerical integration of differential equations.

Under the scenario of only extant tip sampling (that is, $\psi = 0$), a range of models has been introduced with macroevolutionary applications in mind. Foundational work includes Morlon, Parsons and Plotkin (2011) modelling time-continuous changes of speciation and extinction rates and Etienne et al. (2011) modelling density-dependent effects.

We showed that a constant birth and constant death rate can be estimated from a tree when $\rho = 1, \psi = 0$. When we have additional unknown parameters (such as rates changing through time or the sampling parameter), it is possible that not all values can be inferred from the tree since different parameter combinations can lead to precisely the same phylodynamic likelihood value (recall that the phylodynamic likelihood is $L(\eta = (\beta, \delta, T, \rho, \psi, r); \mathcal{T}) = P(\mathcal{T}|\eta)$ with $P(\mathcal{T}|\eta)$ derived in Theorem 9.1.7). In other words, the likelihood surface could have a ridge.

Such a ridge exists under a model with the three free parameters β, δ, ρ (while $\psi = 0$). These three parameters always appear as $\beta\rho$ and $\beta - \delta$ in the terms $\beta p(t|1)$ and $\frac{p(t|1)}{1-p(0|1)}$ (Stadler 2013) which give rise to the likelihood (Theorem 9.1.7). This implies that if a $\hat{\beta}, \hat{\mu}$ maximises the likelihood for $\rho = 1$, parameter combinations like $\rho = 0.5$ and $\tilde{\beta} = 2\hat{\beta}$ and $\tilde{\delta} = 2\hat{\beta} - (\hat{\beta} - \hat{\delta})$ would result in the same likelihood value; and this property holds for any tree. Thus, only two out of the three birth-death parameters are identifiable (Stadler 2009), even in the case of infinite data; infinite data here means infinitely many trees of age T for which we simultaneously estimate their shared birth, death, and sampling rates, with the full phylodynamic likelihood being the product of the phylodynamic likelihood for each tree[1]. Similarly, if sampling through time is modelled with $\psi > 0$ (with $r = 1$ and $\rho = 0$), the three parameters β, δ, ψ always appear as $\beta\psi$ and $\beta - \delta - \psi$ in the likelihood function meaning only two out of the three parameters are identifiable (Stadler et al. 2011; Stadler et al. 2013). Note that if one of the three parameters is known, the other two parameters can be estimated based on a tree using the maximum likelihood technique.

These non-identifiability results were recently generalised for time-varying birth-death models (Louca and Pennell 2020; Louca et al. 2021). They show that for a fixed $\psi = 0$, assuming that birth and death rates are allowed to change arbitrarily through time, the birth and death rate trajectories cannot both be inferred even when the sampling parameter ρ is known. Instead, very different birth and death rate trajectories can explain the tree equally well. A set of trajectories that produce the same phylodynamic likelihood value for all trees is called a *congruence class*. Significantly, this result does not depend on the amount of data used for inference: congruent trajectories remain impossible to distinguish even with infinite data (meaning infinitely many trees, see above). In practice, congruent trajectories may be very different, making it impossible to distinguish, for instance, if rates increase, decrease, or remain constant over time

[1] While in phylodynamics, infinite amount of data correspond to infinitely many trees, in phylogenetics, infinite amount of data correspond to infinitely long sequences (Chapter 6).

in a clade. However, Legried and Terhorst (2022) and Legried and Terhorst (2023) showed that models with rates specified as piecewise-polynomial intervals are identifiable — meaning the birth and death rate trajectories can be inferred — as long as the sampling parameter is provided. This means that subtly restricting the shape of the birth and death rate trajectories will make the model identifiable. Additionally, Truman et al. (2024) have recently shown that models that generate sampled ancestors are always identifiable (this means the removal parameter r is less than 1). Furthermore, Morlon, Robin and Hartig (2022) discuss how the identifiability results of Louca and Pennell (2020) impact birth and death rate estimation in general and highlight that phylogenies remain a valuable source of information for birth and death rates when used, for example, within a hypothesis-driven framework (that is when comparing a small number of plausible models), using parsimony principles or the so-called regularisation techniques (that favour the "simplest" model if several are plausible), as well as by adding non-phylogenetic data (such as fossils or classic epidemiological data).

We conclude this part by highlighting that all co-existing lineages have the same birth, death, and sampling parameters in the models discussed above. Such "neutral" models were generalised to *multi-type birth-death models* where different co-existing lineages of the phylogeny follow different birth and death rates (Maddison, Midford and Otto 2007; Stadler and Bonhoeffer 2013; Kühnert et al. 2016). To calculate the phylodynamic likelihood under such models, we need to formulate separate differential equations for each state. Under multi-type models, we no longer obtain a uniform probability on ranked labelled tree topologies; thus, the tree topology and the branching times together inform the birth and death parameters. We will discuss these models in more detail in Section 9.5.

9.1.6 Applications

9.1.6.1 Epidemiology: quantifying the spread of Ebola in the West African epidemic of 2013-2016

In epidemiology, a quantity of interest is the *basic reproductive number,* R_0 (Anderson and May 1979). R_0 is the expected number of secondary infections caused by a single infected individual introduced into an entirely susceptible population. The value of this parameter is a strong indicator of the fate of an epidemic: if $R_0 < 1$, the epidemic will eventually die out, whereas if $R_0 > 1$, then the infected population size will increase on average, and the epidemic will spread. Furthermore, the value of R_0 indicates the amount of public health effort required to contain an epidemic outbreak. If we assume the constant rate birth-death model, the basic reproductive number can be calculated as $R_0 = {}^{\beta}/_{(\delta+r\psi)}$ (Gavryushkina et al. 2014) (for a more general overview on modelling epidemics, see e.g. Keeling and Rohani (2008)).

In what follows, we calculate R_0 for the West African Ebola outbreak from 2013 to 2016. 72 Ebola genomes from different patients in a Sierra Leone outbreak were published in August 2014 (Gire et al. 2014). The reconstructed phylogenetic tree of the samples from Sierra Leone and 3 sequences from Guinea is shown in Gire et al. (2014, Figure 3B). Here, we estimate R_0

based on this tree. As these genomes were sampled early in the epidemic, we assume a constant transmission rate, β, and constant rate of becoming uninfectious (that is, rate of recovery or death, with or without sampling), $\delta + r\psi$, with $r = 1$. The parameters β and δ correspond to the birth and death rates in the constant rate birth-death model. No samples from the present were available, so we set the extant tip sampling probability $\rho = 0$. Using estimates from other studies, we set the sampling probability prior to the present to ${}^{\psi}/_{(\delta+\psi)} = 0.7$ (meaning $\psi = {}^{7}/_{3}\delta$).

We obtain the maximum likelihood estimates $\hat{\beta}$ and $(\delta \hat{+} \psi)$ using the phylodynamic likelihood, and through this we calculate $R_0 = {}^{\hat{\beta}}/_{(\delta \hat{+} \psi)} = 1.34$ with the confidence interval $CI = [1.12, 1.55]$ (Stadler et al. 2014).

In this example, we used a fixed tree from Gire et al. (2014) and therefore ignored any uncertainty in the tree. Phylogenetic trees obtained from pathogen sequences from an epidemic outbreak often show high uncertainty in the estimated tree (see Section 7.4.5 for assessing uncertainty). In Chapter 10, we will see how Bayesian methods can provide phylodynamic parameter estimates (such as estimates for β, δ, ψ, R_0) by accounting for this phylogenetic uncertainty. Section 10.3 will discuss an Ebola analysis using such Bayesian methods.

9.1.6.2 Epidemiology: quantifying the basic reproductive number of SARS-CoV-2 at the start of the pandemic

In the first few months of 2020, genomes were becoming increasingly available. Thus, many scientists in the field of phylodynamics began to tentatively apply the methods described in this chapter to characterise the developing situation.

One early analysis, performed by some of us, used the birth-death model to infer country-dependent R_0 values for China, Italy, Washington State (USA), and the Diamond Princess cruise ship. Sequences for a location were selected from public sources. These sequences were from samples belonging to patients at the specific location. Additionally, the sequences were from patients in other countries who had recently travelled from the location of interest, meaning returning travellers were considered as a sentinel population, giving insight into the transmission dynamics at the location of interest.

As in the Ebola virus analysis, we assumed a birth-death model in which individuals are sampled through time with rate ψ (and $r = 1$) and noted that $R_0 = {}^{\beta}/_{(\delta+\psi)}$. We fixed the sum of the sampling and death rates (the "become uninfectious" rate) to 36.5/lineage/year, corresponding to an average infectious period of 10 days. Similarly, we assumed that the birth rate β was constant throughout the analysis as we were studying only a short period at the start of the epidemic, before the introduction of major public health measures. The sole exception to this rule was the Diamond Princess, where a ship-wide quarantine was imposed immediately upon detection of the first case and thus before any of the samples corresponding to the sequences were collected. Thus, for the Diamond Princess, we included both before-quarantine and after-quarantine birth rate parameters and used only the first to calculate R_0.

We assumed the HKY substitution model (Section 5.3.4) with two different mean substitution rates (5×10^{-4} or 1×10^{-3} substitutions per site per year), chosen to explore how strongly the results depend on this parameter (for which there was some uncertainty at the time).

Analyses were carried out separately for each location-specific outbreak and used to infer the R_0 parameters. Conceptually, one can view these analyses as a two-step process: First, inferring the tree under the substitution model, then inferring the birth-death model parameters from the tree. In reality, these analyses were conducted jointly using the Bayesian techniques, which we will discuss in Chapter 10.

Figure 9.12 shows the results of the four analyses, repeated for each of the two mean substitution rates. The distributions show the support for R_0 values for each population. In the analyses under the slower substitution rate, the R_0 estimates are centred around 3 for the country-based outbreaks, with a slightly higher value for the ship-based outbreak. For the faster substitution rate, estimates are slightly higher — particularly for China. Those analyses were performed in real-time, and results were shared via a forum on March 1, 2020 (Vaughan et al. 2020).

The study was eventually extended to include other countries and more sequence data, as is described by Vaughan et al. (2024). Figure 9.13 shows the results of these analyses, supporting an R_0 of SARS-CoV-2 of well above 2 — which led to the rapid global spread we had to experience in early 2020.

9.1.6.3 Macroevolution: estimating diversification rate changes through time in mammals

About 65 million years ago, a meteorite hit Earth and caused mass extinction, specifically the extinction of dinosaurs. Palaeontological data led to the hypothesis that this event was followed by a period of increased mammalian diversification (Archibald and Deutschman 2001), with the diversification rate defined as $d =$ (speciation rate − extinction rate).

In what follows, we investigate whether the phylogenetic tree of mammals also supports increased diversification of mammals after the extinction of dinosaurs (Stadler 2011). We use the phylogenetic tree of extant mammals from Bininda-Emonds et al. (2007), shown in Figure 1 of the article. An extended birth-death model was applied to this tree, allowing the parameters to change piece-wise through time (with $\psi = 0$ and $\rho = 1$). That is, rates are constant until time t_1, then change to other constant values until time t_2, and so on.

The results, presented in Figure 9.14 (blue), show that the maximum likelihood diversification rate $\hat{d}$ (which is birth rate minus death rate) was roughly 0.05 until 35 million years ago, where there was a peak followed by a decline in diversification rate[2]. As shown in Figure 9.14, the

[2] Note that the peak observed 35 million years ago may be a flaw of the analysis. In Bininda-Emonds et al. (2007), the authors pulled together unresolved lineages that potentially led to too many diversification events around 35 million years ago, leading to an overestimation of diversification around that time point.

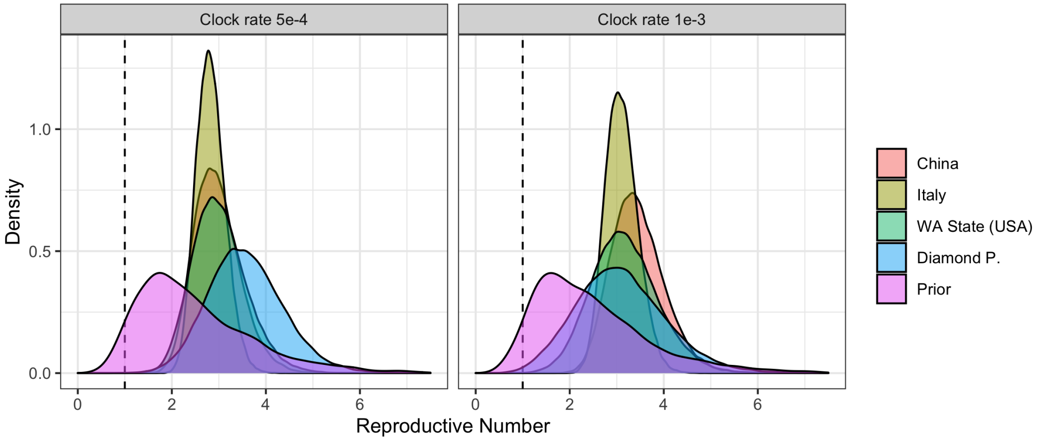


Figure 9.12: Phylodynamic estimation of the basic reproductive number R_0 of SARS-CoV-2 for individual outbreaks, compared with the prior. In the case of the Diamond Princess, this estimate corresponds to the period before the implementation of a quarantine. R_0 is obtained by dividing the birth rate by the death and sampling rate. We show distributions rather than point estimates, as the analysis was performed within the Bayesian framework (Chapter 10). The dashed vertical line shows an R_0 of 1. This analysis was performed as the outbreaks unfolded and was shared on a forum on March 1, 2020. Figure adapted from Vaughan et al. (2020).

analysis shows no evidence of an increase or decrease in diversification rate around 65 million years ago.

Stadler (2011) assessed the uncertainty in the estimate of the diversification rate using a parametric bootstrapping approach (Section 7.4.4). Multiple trees were simulated using the estimated maximum likelihood parameters. Based on these simulated trees, birth-death parameters were re-estimated from the simulated phylogenies. Re-estimated maximum likelihood diversification rates are consistent with the original estimate, as shown in Figure 9.14 (black).

Phylogenetic analysis based on a different mammalian phylogenetic tree (Meredith et al. 2011) did not show evidence for an increase or decrease in diversification rate around 65 million years ago either. Why does the phylogenetic inference disagree with the hypothesis based on paleontological data? This is still an open area of research, and we see the potential to shed light on this question by jointly analysing fossil and phylogenetic tree data (Zhang et al. 2016; Gavryushkina et al. 2017).

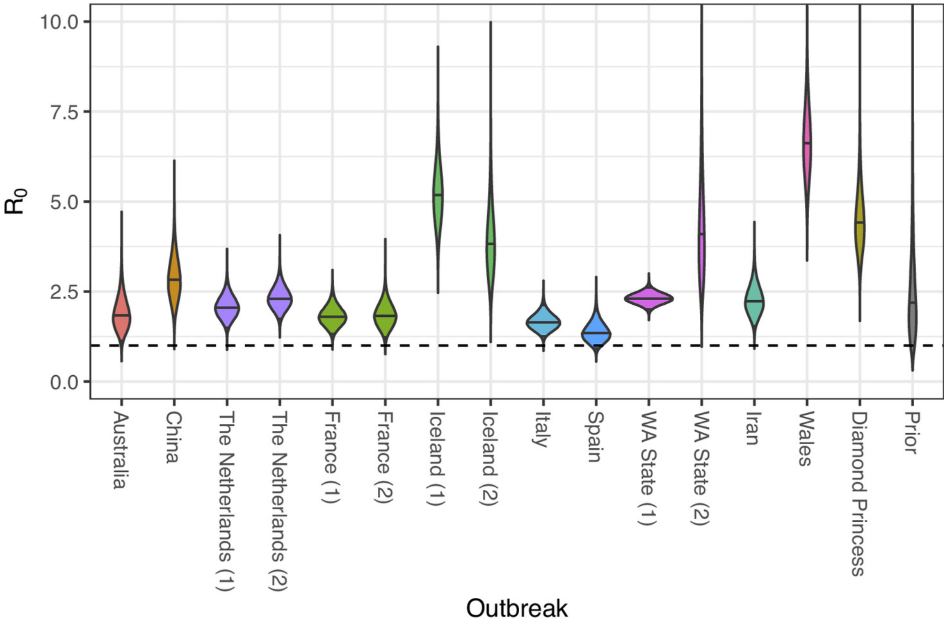


Figure 9.13: Phylodynamic estimation of the basic reproductive number R_0 of SARS-CoV-2 in different populations based on sequence data. The real-time analyses performed for Figure 9.12 were expanded and published in a peer-reviewed journal retrospectively; see Vaughan et al. (2024) for details. Figure adapted from Vaughan et al. (2024).

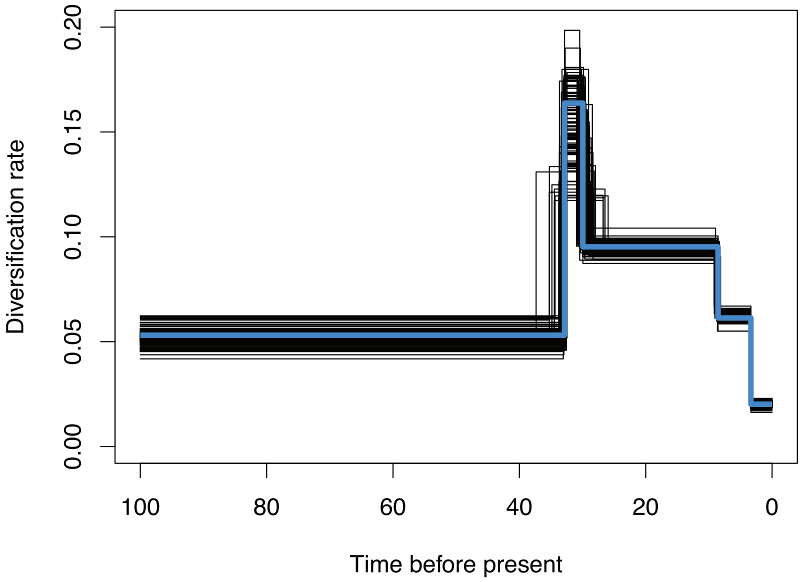


Figure 9.14: Maximum likelihood diversification rate estimate through time (blue) based on a mammal phylogeny. The black lines indicate the parametric bootstrap interval, which was obtained by simulating birth-death trees using the maximum likelihood parameters, and then re-estimating the diversification rate of each simulated tree (black). No signal is visible for elevated diversification upon dinosaur extinction around 65 million years ago. Figure adapted from Stadler (2011).

9.2 Coalescent theory

We now introduce the *coalescent model*, or simply the coalescent, as a modelling framework for trees. The original coalescent model was introduced by John Kingman (1982) as a way to model allele frequency dynamics; it is the basis of many studies in population genetics. More recently, this model and its extensions have been widely applied to estimate processes within the field of population dynamics, using the coalescent within phylodynamic inference. In this context, coalescent models naturally allow the population size to become an explicit target of phylodynamic inference. At its core, the coalescent is a backwards-in-time process that starts with extant lineages at present (tips of the tree) that coalesce to give rise to a tree. Recall that the birth-death models described in the previous section are forwards-in-time processes starting from the first individual and ending at present; we compare the two frameworks in more detail in Section 9.3.

9.2.1 The Wright-Fisher process

To develop a quantitative connection between population size and phylogeny, we first need to define a model for propagating heritable traits within a fixed-size population. One such model is the *Wright-Fisher process* that models genetic drift (Wright 1931; Fisher 1930), which has been a cornerstone of population genetics since its introduction by the founders of the field.

In the classic Wright-Fisher process, we assume a population with a constant size of N individuals. While these individuals may differ in genetic make-up or other ways, the model itself is completely blind to these differences, treating every individual equally.

The model assumes discrete, non-overlapping generations. Each member of a given generation has exactly one parent in the previous generation. Thus, the model is most directly applicable to biological units undergoing asexual reproduction (see Section 1.1.1). However, these units might be genes or other genetic elements that can be treated as reproducing asexually, although they belong to sexually reproducing organisms.

The choice of a parent from the previous generation is completely random. Importantly, this means that the choice is independent of any trait associated with the individual or its parent. It is, therefore, a neutral model, as any selection of parents based on fitness criteria is not allowed. Correspondingly, each parent may have zero, one, or many children in the next generation, but the total number of children in a generation must equal the population size N.

Figure 9.15 represents the basis on which the Wright-Fisher model operates: the number of individual population members in each generation through time. Figure 9.16 shows a particular realisation of the Wright-Fisher model, which results in assigning children to parents between generations of the population. The thick lines and the filled circles represent the phylogenetic relationships between three arbitrarily chosen (or sampled) individuals. These

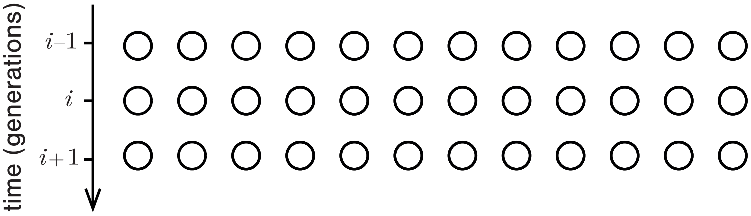


Figure 9.15: The Wright-Fisher model is a population genetic model based on a sequence of discrete generations of a population of a constant size. For example, the above schematic represents three generations of such a population, where each circle represents an individual member of the population, and each row of circles represents the members of a single generation. Time is measured in generations from past to present, shown on the y-axis.

relationships are implied by the particular outcome of the Wright-Fisher process and define the phylogenetic tree for the three sampled individuals.

9.2.1.1 Most recent ancestor of two samples

In the simplest case, when two samples are drawn randomly from the present population, quantifying their phylogenetic tree is reduced to the following question: what is the probability that the *most recent common ancestor (MRCA)* of these two samples occurred m generations before the present?

To answer this, consider that:

(i) since each individual picks their parent uniformly at random, the probability that two individuals from the same generation have the same parent is $1/N$;

(ii) the probability that two individuals in the same generation do not have a common ancestor in the previous generation is $(1 - 1/N)$.

Thus, the probability for the two sampled individuals to share a common ancestor in the mth generation before the present is the product of the probability of no common ancestor in the first $m-1$ generations and the probability of a common ancestor in the mth generation before the present:

$$P_{\mathrm{MRCA}}(m) = (1 - \frac{1}{N})^{m-1}\frac{1}{N}. \tag{9.50}$$

This is simply a geometric distribution (Box 15 on page 86) with a success probability of $1/N$. Since the mean of such a distribution is equal to the inverse of the success probability, we must wait on average N generations to see a common ancestor of two samples from a population of size N.

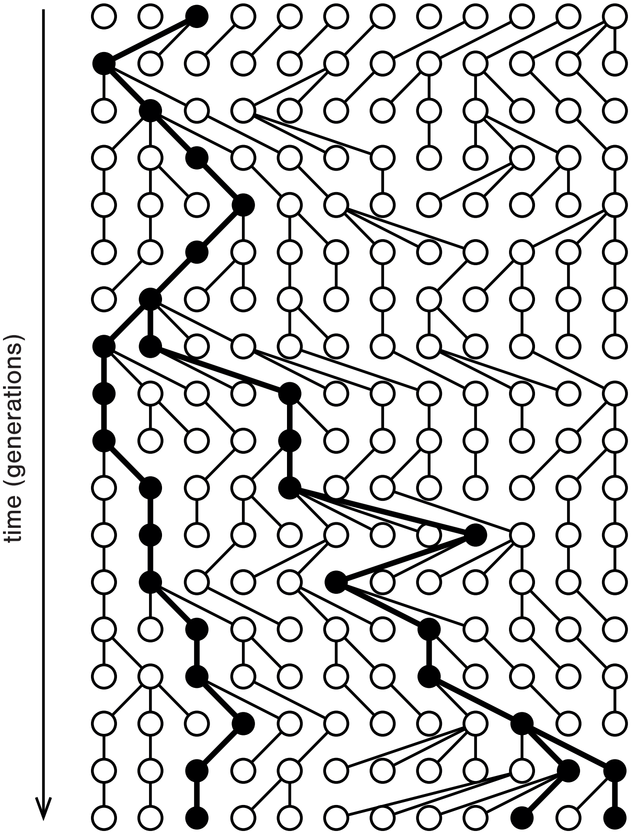


Figure 9.16: A realisation of the Wright-Fisher model across multiple generations. As in Figure 9.15, each row of circles represents a single generation, while time, measured in generations, increases downwards. Lines between the circles represent the randomly chosen parent-child relationships; each individual chooses a parent uniformly at random from all individuals in the previous generation. The three filled circles in the last generation (bottom row) represent three arbitrarily chosen sampled individuals. The thick lines and filled circles in earlier generations represent the ancestry of those three sampled individuals obtained by "tracing back" along the parent-child relationships. Note that while the lines representing parent-child relationships may cross, it is always possible to reorder the members of each row so that these lines do not cross — as we have done here.

This result can be extended to develop the full probability of a larger sampled tree under the discrete-time Wright-Fisher model. In such a tree, internal nodes would occur at integer generation numbers and could involve more than two child lineages, producing non-binary trees (see also Figure 9.16). Instead, at this point, we will leave the discrete-time Wright-Fisher model and begin to develop an approximate coalescent model for continuous-time binary phylogenetic trees.

9.2.2 Kingman's coalescent process

Kingman's coalescent (also known as Kingman's n-coalescent, where n refers to the number of sampled lineages at the start of the process) is a continuous-time Markov chain (see Box 24 on page 98) which produces time trees. The process runs backwards in time, building the tree through successive pairwise merging events known as coalescence events or simply coalescences. There are several distinct population genetic models for which Kingman's coalescent arises as a limiting case. Here, we introduce the coalescent using the Wright-Fisher model that was discussed in the previous section.

9.2.2.1 Coalescence rate between two lineages

We consider a pair of individuals sampled from a particular generation of a Wright-Fisher population. In the previous section, we derived that the probability that these individuals share a parent in the previous generation is $1/N$ and that the number of generations until they share a common ancestor follows a geometric distribution. When developing the corresponding Kingman coalescent model, we consider what happens when N is very large, meaning that the probability of a pair of lineages finding a common ancestor in any given generation becomes extremely small. As discussed in Box 18 on page 90, this is the limit at which a geometric distribution approaches an exponential distribution.

More specifically, the geometric distribution (Equation (9.50)) for the pair of lineages coalescing under the Wright-Fisher model has a success probability of $p = 1/N$. Using g to represent the time interval between successive generations, and $t_2 = mg$ as the time of the mth generation before the present, we define the coalescence rate $\theta = p/g = 1/Ng$, that is, the probability of a pair of lineages coalescing per unit of time.

With these definitions, we can apply the limit described in Box 18 on page 90 to derive the probability density of observing a coalescence at time t_2 before the present, in the limit of $g \to 0$ and $N \to \infty$, where $\theta = 1/Ng$ is constant:

$$f(t_2|\theta) = e^{-\theta t_2}\theta. \tag{9.51}$$

In this limit, referred to hereafter as the *coalescent limit*, the time until a pair of lineages coalesce is thus exponentially distributed with rate θ.

9.2.2.2 Coalescence rate between more than two lineages

This naturally leads to the probability density for the first coalescence between any pair selected from k lineages per unit time, where $k \geq 2$. Importantly, in our large N limit, the probability of more than two lineages finding a common ancestor simultaneously approaches zero. This means that the time of the first coalescence event is simply the minimum of the $\binom{k}{2}$ pairwise coalescent times, each exponentially distributed with rate θ. We can then use the result given in Box 17 on page 89 to find that the minimum is itself exponentially distributed with a rate equal to the sum of the pairwise rates, that is,

$$f(t_k|\theta) = \exp\left(-t_k \binom{k}{2} \theta\right) \times \binom{k}{2} \theta. \tag{9.52}$$

As usual, we can interpret this as the product between the probability of no coalescence occurring in time t_k (the exponential function) and the probability density of a coalescence occurring immediately after.

This probability density was first calculated by John Kingman (1982).

9.2.2.3 The coalescent process and the probability density of a coalescent tree

So far, we have derived the probability distributions for the time taken for different numbers of sampled lineages to coalesce by considering the limit of the Wright-Fisher process when N goes to infinity. This is all required to define the *coalescent model*: a stochastic process that produces sampled coalescent trees. This process is a continuous-time Markov chain (Box 24 on page 98) on the sampled lineages at a particular time. It starts at present with the individuals sampled at present, moving into the past, producing coalescence events governed by the probability density in Equation (9.52). Each event merges a randomly chosen pair of sampled lineages (lineages with a sampled descendant) into a new internal tree node, reducing the number of ancestral lineages by 1. The Markov chain process terminates when a single lineage remains.

The probability density of a labelled tree generated by this process can be expressed as a product between the probability for the time intervals between coalescence events, the probability densities (the rates) of coalescences occurring at times corresponding to the internal nodes, and the probability of the particular pair coalescing at that node. For example, the

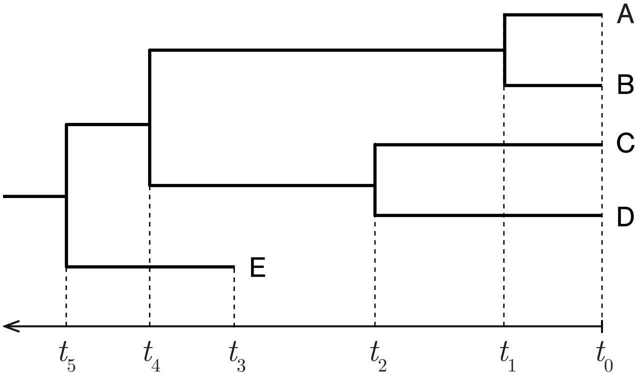


Figure 9.17: A labelled tree with coalescence and sampling event times.

probability density for the labelled tree given in Figure 9.17 can be written as

$$\begin{aligned} f(\mathcal{T}|\theta) = {} & \exp\left(-(t_1 - t_0)\binom{4}{2}\theta\right)\theta \\ & \times \exp\left(-(t_2 - t_1)\binom{3}{2}\theta\right)\theta \\ & \times \exp\left(-(t_3 - t_2)\binom{2}{2}\theta\right) \\ & \times \exp\left(-(t_4 - t_3)\binom{3}{2}\theta\right)\theta \\ & \times \exp\left(-(t_5 - t_4)\binom{2}{2}\theta\right)\theta. \end{aligned} \tag{9.53}$$

Notably, in the above expression, the binomial coefficients appear only in the exponential functions representing the probability of no coalescence in each interval, but do not appear in the factors to the right of the exponentials. This is because we derive the probability density of a labelled tree, that is, we distinguish between coalescences involving different pairs of lineages: the probability of a chosen pair of k lineages coalescing is less than the probability of any arbitrary pair coalescing, by a factor of $\binom{k}{2}$.

Additionally, the third line is missing the coalescence rate term entirely. This is because, for this particular tree, t_3 corresponds to a sampling event. The coalescent process conditions on such events explicitly; thus, the event does not contribute to a rate term but merely increases the number of lineages by 1.

In general, the probability density for a labelled tree $\mathcal{T}$ under the coalescent can be expressed in terms of intervals between consecutive coalescence or sampling events. For a tree with n leaves, we have $n - 1$ coalescent times. Furthermore, assume the leaves are sampled at m

different times (in Figure 9.17, we have $m = 2$). In total, we then have $m + n - 1$ coalescence or sampling events that happened at $m + n - 1$ time points. Defining t_i as the time of event i, k_i as the number of ancestral lineages extant in the interval between t_i and t_{i-1}, and ν_i as 1 if the event i is a coalescence event and 0 otherwise, we can state the following:

Theorem 9.2.1. *The probability density for a labelled tree $\mathcal{T}$ with n leaves sampled at m times under the coalescent is given by:*

$$f(\mathcal{T}|\theta) = \prod_{i=1}^{m+n-1} \exp\left(-(t_i - t_{i-1})\binom{k_i}{2}\theta\right)\theta^{\nu_i}. \tag{9.54}$$

Interpreting this expression as a function of θ (see Box 25 on page 116) is the likelihood function for the coalescence rate parameter θ given the tree $\mathcal{T}$. This implies that we can infer θ based on a phylogenetic tree, assuming the conditions of the coalescent model are met.

Note that given all n samples are collected at present, the probability density is a function of the branching times $t_1 < t_2 < \ldots t_{n-1}$ in the tree and the coalescent parameter θ. Specifically, the probability density of the tree is independent of the tree topology, as was also the case under the birth-death model (Section 9.1.3). Analogously to the proof in Theorem 9.1.3, we can show that each ranked labelled tree on n tips has the same probability under the coalescent (this also holds for all the extensions of the coalescent model without structure in the population; for structured models see Section 9.5).

9.2.2.4 The expected height of a coalescent tree

An interesting consequence of the coalescent process is that the expected time required for n lineages to coalesce into 1, that is, the expected age of a coalescent tree with n leaves sampled at present is

$$\begin{aligned}
E(t_{root}) &= \sum_{k=2}^{n} \frac{Ng}{\binom{k}{2}} \\
&= Ng\sum_{k=2}^{n} \frac{2}{k(k-1)} \\
&= 2Ng\sum_{k=2}^{n}\left(\frac{1}{k-1} - \frac{1}{k}\right) \\
&= 2Ng\left(\sum_{k=1}^{n}\frac{1}{k} - \sum_{k=2}^{n}\frac{1}{k}\right) \\
&= 2Ng(1 - \frac{1}{n}),
\end{aligned} \tag{9.55}$$

which approaches the upper bound of $2Ng$ as the number of samples increases.

Of course, when leaf nodes are sampled through time, the root may be arbitrarily old compared to the most recent sample.

9.2.2.5 Coalescence rates and finite population sizes

Before proceeding, let us take a moment to consider what these results imply from a practical perspective. Strictly speaking, the probability density above only exists in the limit of infinite population sizes ($N \to \infty$) and zero generation times ($g \to 0$). However, any real population with discrete generations will have a finite number of individuals and non-zero times between generations. Does this mean that the distribution above, and, by extension, the coalescent model, is useless?

Not at all! While the probability density for the time between coalescence events is only exactly exponential in the specified limit, it remains a good approximation for the corresponding probability distribution in finite populations with large population sizes and relatively short generation times. In such cases, we interpret the coalescence rate θ as approximately equal to $^1/_{Ng}$, where N is the population size and g is the inter-generation time for the finite population.

From this point on, we will always write $^1/_{Ng}$ instead of θ, keeping in mind that the expressions are, strictly speaking, only approximations when written this way. This reflects the way coalescent models are used in phylodynamics.

9.2.3 Population dynamics

So far, we have considered a constant population size N. However, the sizes of real populations usually change over time. An obvious extension is to replace this constant with a time-dependent population size function $N(t)$. Throughout this section, we explain how to quantify population dynamics expressed as $N(t)$ based on a phylogenetic tree.

For example, we might define an exponentially growing population $N(t) = e^{-rt}N_{\text{present}}$, where r is the growth rate and N_{present} is the population size at present. Here, the minus sign in the exponential is because t increases backwards in time. As we will outline below, these population size changes over time will impact the shape of the sampled trees.

Incorporating time dynamics into the coalescent is straightforward. For instance, Griffiths and Tavaré (1994) showed that in the coalescent limit, the coalescence rate between a pair of lineages in a Wright-Fisher model with time-dependent population size is $^1/_{N(t)g}$. Here, $N(t)$ is the continuous-time large population size limit of the discrete-time Wright-Fisher population

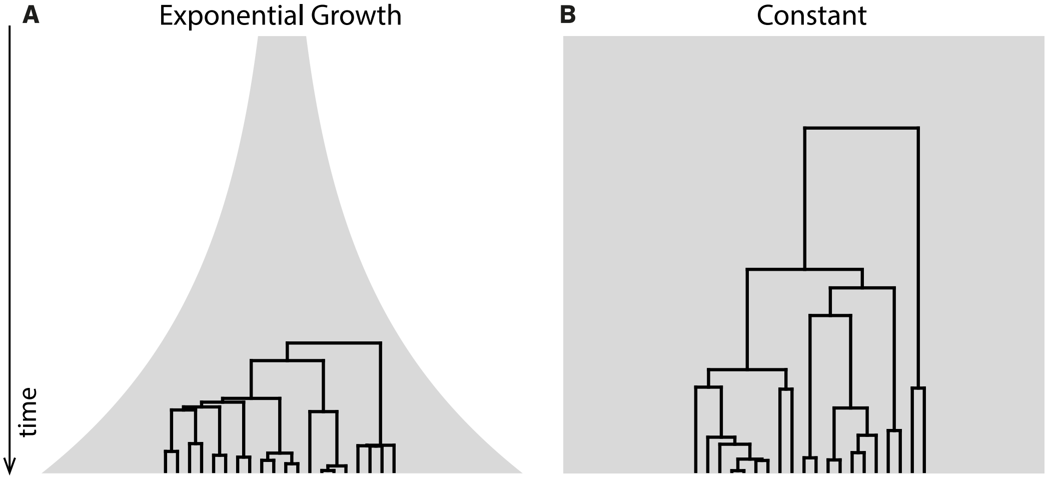


Figure 9.18: Trees generated by coalescent processes assuming different population dynamics. **A** shows a tree generated by a coalescent, assuming an exponential population growth model. **B** shows a tree generated by a coalescent assuming a constant population size model. The width of the grey background indicates the population size at different times.

function. This means that the coalescence rate is larger when the population is small than during periods when it is large. The probability of a labelled tree becomes

$$f(\mathcal{T}|N(t)g) = \prod_{i=1}^{m+n-1} \exp\left(-\binom{k_i}{2}\int_{t_{i-1}}^{t_i} \frac{\mathrm{d}t}{N(t)g}\right)\left(\frac{1}{N(t_i)g}\right)^{\nu_i}, \tag{9.56}$$

where k_i is the number of ancestral lineages extant in the interval between t_i and t_{i-1}. Note that for constant N, this formula simplifies to Theorem 9.2.1. An example of the effect of population size variation on the shapes of trees is shown in Figure 9.18. The larger the population size, the larger the waiting time until a coalescence event since the continuous-time rate of coalescence $^1/_{N(t)g}$ will be smaller. We can compare an exponentially growing population of size $N_1(t)$ with a population of constant size N_2, where both have the same present-day population size $N_1(0) = N_2$. All coalescence rates in the exponentially growing population will be larger than or equal to the coalescence rates in the population of constant size, leading to shorter trees in the exponential scenario. Consequently, the timing of coalescence events in phylogenies reconstructed from the sampled sequences can inform us about the total population size changes over time.

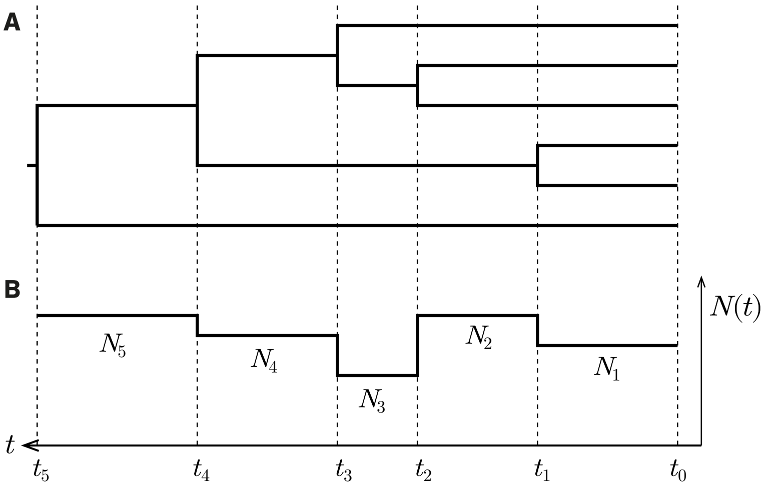


Figure 9.19: **A** Tree and a corresponding **B** skyline plot. The skyline plot quantifies population dynamics, allowing each interval between coalescence events to have a distinct population size.

9.2.3.1 Non-parametric inference of population dynamics

What if we do not know (or do not want to assume) that population size is governed by a particular parametric model (a model with a finite number of parameters, such as exponential growth)? In this case, we can use the so-called *non-parametric methods*, which use models where the number of free parameters grows with the number of samples (see also Section 7.4.3).

The most well-known example is the *skyline plot* developed by Pybus, Rambaut and Harvey (2000). In this model, the probability of a tree with all samples at present is given by[3]

$$f(\mathcal{T}|\vec{N}g) = \prod_{i=1}^{n-1} \exp\left(-(t_i - t_{i-1})\binom{k_i}{2}\frac{1}{N_i g}\right)\frac{1}{N_i g}, \tag{9.57}$$

where $\vec{N} = (N_1, \ldots, N_{n-1})$ is a vector of length $n-1$. Thus, this vector has as many elements as coalescence events in the tree. This model assumes that population size is constant within each time interval between merging events, as illustrated in Figure 9.19.

As in the other models, the expression for the tree probability density function $f(\mathcal{T}|\vec{N}g)$ is the likelihood $L(\vec{N}g|\mathcal{T})$ for the elements of $\vec{N}g$ given a tree. Given a tree, we can calculate a

[3] The original model described by Pybus, Rambaut and Harvey (2000) was expressed slightly differently but is equivalent to what we present here.

maximum likelihood estimate of the population dynamics, in this case, the population sizes through time.

In fact, we can derive a closed-form expression for the maximum likelihood estimates. To do this, note that the complete likelihood for $\vec{N}g$ can be written as the product of the likelihoods of the individual elements:

$$L(\vec{N}g|T) = \prod_{i=1}^{n-1} L(N_i g), \tag{9.58}$$

where

$$L(N_i g) = \exp\left(-(t_i - t_{i-1})\binom{k_i}{2}\frac{1}{N_i g}\right)\frac{1}{N_i g}. \tag{9.59}$$

Then we define $\hat{N}_i$ as the maximum likelihood estimate of N_i, which by definition satisfies

$$\frac{\mathrm{d}}{\mathrm{d}N_i} L(N_i g)\bigg|_{N_i=\hat{N}_i} = 0. \tag{9.60}$$

Since $\log(x)$ is a function monotonically increasing in x, we can apply the same optimality condition to $\log L(N_i)$ and optimise with respect to N_i^{-1}, getting

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}N_i^{-1}} \log L(N_i g)\bigg|_{N_i=\hat{N}_i} &= 0 \\
&= \frac{\mathrm{d}}{\mathrm{d}N_i}\left(-(t_i - t_{i-1})\binom{k_i}{2}\frac{1}{N_i g} + \log(\frac{1}{N_i g})\right)\bigg|_{N_i=\hat{N}_i} \\
&= -(t_i - t_{i-1})\binom{k_i}{2}\frac{1}{g} + N_i\bigg|_{N_i=\hat{N}_i} \\
&= -(t_i - t_{i-1})\binom{k_i}{2}\frac{1}{g} + \hat{N}_i.
\end{aligned} \tag{9.61}$$

Thus, in each interval, we have the following maximum likelihood estimate:

$$\hat{N}_i g = (t_i - t_{i-1})\binom{k_i}{2}. \tag{9.62}$$

While this is the simplest case, extensions to the classical skyline plot method involve allowing the grouping of multiple intervals together (Strimmer and Pybus 2001). Many ways have also been developed to incorporate uncertainty into the results (Drummond et al. 2005; Heled and Drummond 2008) within a Bayesian framework (see also Chapter 10).

9.2.4 Coalescent approximation of birth-death models

As mentioned earlier, coalescent theory is not intrinsically tied to the Wright-Fisher population dynamics model. Indeed, a much broader class of population genetic models has a limit in which the probability of a sampled tree is given by the coalescent process. One of the most important features of population models that possess coalescent limits is the exchangeability of individuals within the population. This requirement forms the basis for the very general Cannings model (Cannings 1974), of which the Wright-Fisher model is a special case. It also encompasses the Moran model (Moran 1962), which is similar to the Wright-Fisher model but which allows overlapping generations. Given that birth-death models described at the beginning of this chapter (see also Theorem 9.1.3) also feature exchangeability of individuals, it is unsurprising that we can approximate these models using a coalescent distribution (Volz et al. 2009; Volz 2012; Volz and Frost 2014).

To see how such an approximation works, consider a typical birth-death trajectory such as the one shown in Figure 9.20. As discussed in Section 9.1.1, under the birth-death model, the birth events occur at a rate βI, where I is the population size. While this exact rate is itself a random variable (since the population size I depends on the outcome of the birth-death process at a given time), we can approximate it using the expected value of the population size under the stochastic model (Section 9.1.1). We thus define the birth rate $B(t) = \beta I(t) = \beta I_0 \exp(-t(\beta - \mu))$ where I_0 is the population size at present.

Compared to the formulation in Section 9.1.1, we changed the sign in the exponent, letting the time t run backwards for consistency with the coalescent. This approximation of the stochastic outcome with the expectation is adequate when $I(t)$ is very large[4].

Consider the two sampled lineages extending from the right-hand side of Figure 9.20. These lineages must coalesce at a time point corresponding to a birth event. As these ancestral lineages propagate backwards in time, every birth event they encounter represents a possible coalescence time. What is the probability that the chosen pair of tree lineages coalesce at a particular birth event? To answer this, consider that all population members are equivalent under our model. The probability that this particular pair of lineages coalesce at a given birth event is the inverse of the total number of such pairs in the population which could coalesce at that time point: $p_2(t) = 1/\binom{I(t)}{2}$. More generally, for k extant lineages in the tree, the probability of a coalescence occurring among them at a given birth event is the ratio of the number of pairs among those k lineages to the total number of pairs in the population. That is, $p_k(t) = \binom{k}{2}/\binom{I(t)}{2}$.

By combining the per-birth-event coalescent probability $p_k(t)$ with the population birth rate

[4] In fact, one can show that the standard deviation in the population size of this approximation vanishes relative to the population size for the simple birth-death model in the large population limit when $\beta > \delta$, meaning the deterministic approximation becomes arbitrarily good in this case.

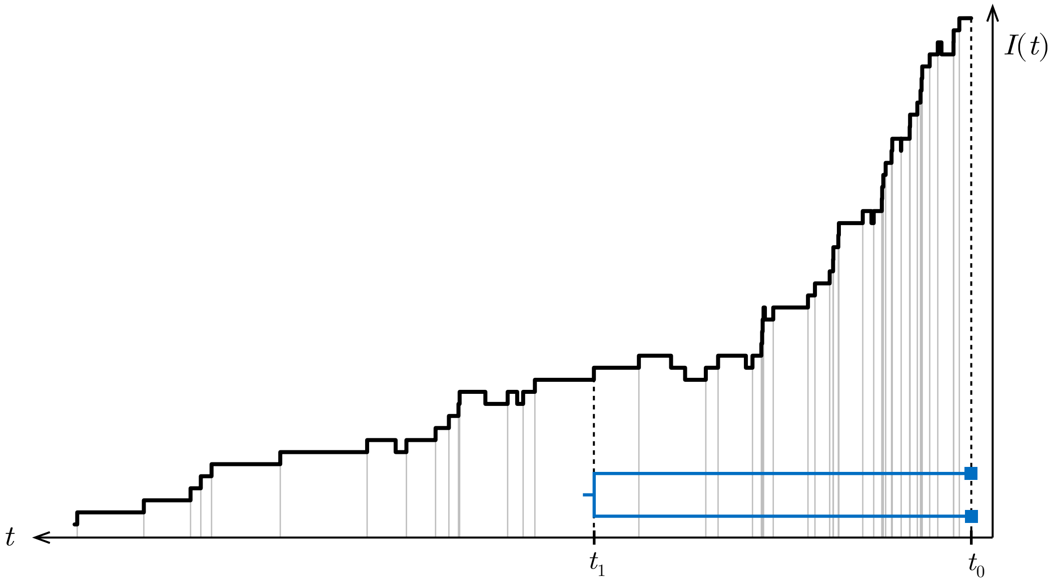


Figure 9.20: Relationship between a birth-death process trajectory (black) and a possible coalescence time of a pair of sampled lineages at time t_1 (blue). The x-axis denotes time, with t_0 being the present, and the y-axis the size of the population $I(t)$. Each birth event time, represented by vertical lines at $t > t_0$, represents a possible coalescence time. This consideration leads to the coalescent approximation of birth-death processes described in the text.

$B(t)$, we recover an approximation for the coalescence rate at a given time:

$$\begin{aligned}\chi_k(t) = B(t)p_k(t) &= \frac{\beta I(t)\binom{k}{2}}{\binom{I(t)}{2}} \\ &= \binom{k}{2}\frac{2\beta}{I(t)-1} \\ &\simeq \binom{k}{2}\frac{2\beta}{I(t)}. \end{aligned} \tag{9.63}$$

We drop the -1 from the denominator in the final line since that value is negligible when $I(t)$ is large.

This coalescence rate can be used to compute an approximate probability for a tree for a given set of birth-death parameters. This approximation is valid only when the population size remains large over the entire timespan of the tree, that is, when the deterministic approximation of the stochastic population size is valid.

Interestingly, this coalescence rate is identical to the coalescence rate between lineages of the Wright-Fisher model for a deterministically varying population size $N(t)$ that we discussed in Section 9.2.3. The only difference is that in the Wright-Fisher case, the rate is proportional to $1/g$ (the inverse of the time between generations), while in the approximate birth-death case, the rate is proportional to 2β.

9.2.5 Effective population size

Of course, real populations evolve in far more complex ways than these simple models describe. For instance, real populations are often structured in some way. For example, consider that the population may exist as several distinct communities within which individuals are more likely to share a recent common ancestor than individuals belonging to separate communities (e.g. birds on different islands). Alternatively, one may apply the Wright-Fisher model not only to haploid populations but to populations of genes belonging to sexually reproducing organisms. Non-random mating in those populations can produce a similar effect as population structure, where pairs of "child" genes may be more or less likely to share a "parent" gene in the previous generation depending on the mating preferences of the organisms in which they exist. Furthermore, while completely ignored in the basic Wright-Fisher model, selection can play an extremely important role in shaping trees and, therefore, sequence diversity.

For these reasons, when population size is estimated assuming an ideal Wright-Fisher model, we refer to the result as the *effective population size* (Wakely 2016). Loosely speaking, it is the size of an idealised Wright-Fisher population model with the same genetic diversity as our actual population. The changes in effective population size may reflect underlying actual population size changes but may also be influenced by the presence of population structure (Pannell 2003), expected generation time (Volz et al. 2009), or selection (Charlesworth 2009). The explicit modelling of structured populations is discussed later in Section 9.5.

9.2.6 Application

9.2.6.1 Epidemiology: Hepatitis C epidemic in Egypt

Hepatitis C The hepatitis C virus (HCV) was first identified in 1989 (Choo et al. 1989; Kuo et al. 1989). Its genome is a single-stranded, 9.6 kilobases long RNA molecule.

The World Health Organisation (WHO) estimates that around 1.5 million people are newly infected with HCV every year, of which about 70% develop a chronic infection (WHO 2023). Chronic HCV infections damage the liver, causing liver cirrhosis and increasing the risk of some types of cancer that affect major organs, including the liver and pancreas.

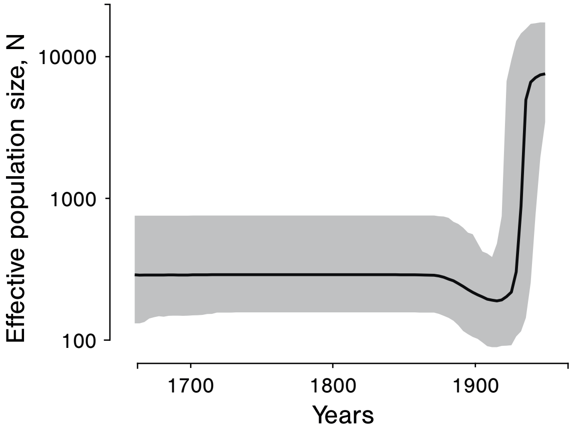


Figure 9.21: Estimate of the infected population size through time obtained from an Egyptian HCV sequence dataset under a coalescent skyline plot. The black line is the median estimate, and the uncertainty interval is shown in grey. The figure was created from the results of the "Skyline plots" Taming the BEAST tutorial `https://taming-the-beast.org/tutorials/Skyline-plots/`.

HCV is mostly transmitted through exposure to infected blood, although other modes, such as sexual transmission and vertical (mother-to-child) transmission, are also possible. Blood transfusions and injections with infected needles account for most new infections.

Hepatitis C in Egypt At the end of the 20th century, Egypt had a Hepatitis C prevalence of around 20%, the highest HCV prevalence in the world at that time (Quinti et al. 1995; Arthur et al. 1997). The neighbouring countries had much lower HCV prevalence. The dominant hypothesis for the high prevalence in Egypt was that HCV was spread by contaminated needles as part of an independent public health campaign against another parasite, namely *Schistosoma* worms (Frank et al. 2000), during the first part of the 20th century. An investigation into the genetic evidence for this explanation formed one of the first applications of coalescent models to phylodynamics (Pybus et al. 2003).

Figure 9.21 illustrates the results of a later analysis by Drummond et al. (2005), based on 63 individual 411 bp HCV E1 gene sequences from Egypt (Ray et al. 2000; Pybus et al. 2003). These sequences were used to infer a phylogenetic tree, which was related to the infected population size through time using the skyline coalescent model introduced in the previous section[5]. The results suggest an almost 100-fold increase in the infected population size during the first half of the 20th century.

[5] In fact, this was done simultaneously via the Bayesian techniques which we will introduce in Chapter 10.

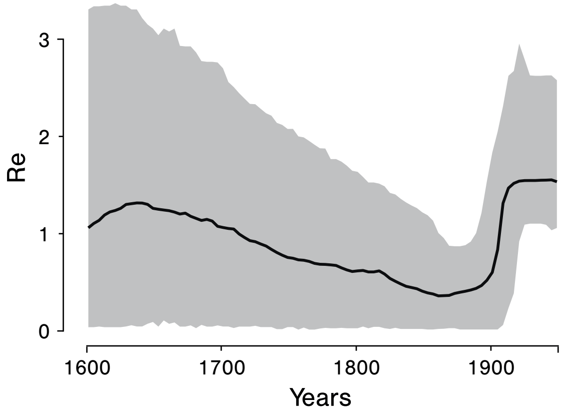


Figure 9.22: Estimate of the reproductive number R_e through time obtained from the same Egyptian HCV dataset as used in Figure 9.21, now assuming a piecewise constant birth-death model (birth-death skyline plot). The effective reproductive number for a particular time point is the birth rate divided by the death rate at that time point. The black line is the median estimate, and the uncertainty interval is shown in grey. The figure was created from the results of the "Skyline plots" Taming the BEAST tutorial `https://taming-the-beast.org/tutorials/Skyline-plots/`.

This result was complemented by another analysis on the same dataset performed eight years later, using the birth-death skyline plot. For each time point, the ratio of estimated birth and death rate was considered, which is the expected number of secondary infections caused by a single infected individual at that time point — called the effective reproductive number R_e[6]. For the HCV analysis, the effective reproductive number estimate also reveals an increase in transmission during the same period as the coalescent analysis, as shown in Figure 9.22 (Stadler et al. 2013).

These estimated timings of the increase in HCV transmission are consistent with the antischistosomal injection therapy explanation for the rapid dissemination of HCV. In particular, injection therapy against *Schistosoma* worms involving potentially contaminated needles was in use from the 1920s through to the 1980s (Frank et al. 2000). Additional detailed analyses show that the estimated effective reproductive number of HCV decreased to 1 when oral therapy for schistosomiasis was introduced (Stadler et al. 2013).

[6] We note that the effective reproductive number is a generalisation of the basic reproductive number. At the start of an outbreak with all individuals being susceptible, the effective reproductive number is, in fact, the basic reproductive number.

9.3 Comparison of coalescent and birth-death models

Both the coalescent and the birth-death models describe $P(\mathcal{T}|\eta)$, the distribution of trees $\mathcal{T}$ given the parameters of the population model η. The coalescent is based on a backwards-in-time process that starts with tips at present, and the branching times in our notation are denoted by $t_1 < t_2 < \dots t_{n-1}$ from present to root, Figure 9.19. It is parameterised by the population size through time and the generation time. That is, it is parametrised by $\eta = (N(t), g)$, which determines the coalescence rate. It is important to emphasise that while the coalescent process for the sampled tree proceeds backwards in time, the population genetic models (such as the Wright-Fisher model) from which the coalescent is derived describe the evolution of the population forwards in time.

In contrast, the birth-death model is a forwards-in-time process that starts with the first individual in the past, and in our notation, the branching times from the root to the present are denoted by $x_1 > x_2 > \dots > x_{n-1}$, Figure 9.10. It is parameterised by the birth and death rates through time together with the start time and sampling parameters, that is, $\eta = (\beta(t), \delta(t), T, \rho, \psi, r)$.

Both models are neutral in the sense that the same dynamics of birth, death, or coalescence rates govern all individuals at a particular time point. Consequently, under both frameworks, each ranked labelled tree on n tips sampled at a single time point has the same probability (Theorem 9.1.3). Interestingly, the critical birth-death process (a birth-death process where the birth rate equals the death rate) and the coalescent also have the same expected branching times (up to a scaling factor that depends on the number of tips), but the branching times have a different variance and different higher moments (Gernhard 2008). In general, the distributions of branching times — including the expectations — differ. These differences stem from differences in modelling assumptions, such as the sampling process, the population size, and generation times.

First, as suggested by the underlying parameters, there is a conceptual difference in how the models deal with samples. Generally, the birth-death models explicitly model the production of samples (we introduced the parameters ρ, ψ, r), while coalescent models generally condition on the given number and times of samples rather than treating these aspects as data. This means that, while analyses based on coalescent methods may not be as easily led astray by a misspecified generative sampling model, birth-death models can use information from the numbers and times at which sequences have been collected to learn more about the population dynamic process. From a simulation perspective, the coalescent model gives rise to trees on some pre-specified sampling times, while for the birth-death model, such times are part of the data, and thus, a simulation just based on the model gives random sampling times. As a consequence, trees from these two models cannot be compared directly (May and Rothfels 2023). We note that approaches to also assume a sampling model under the coalescent were introduced but thus far are not frequently being used (Volz and Frost 2014).

Second, coalescent models generally assume a deterministically varying population[7], while

[7] But see, for example, Popinga et al. (2015), which explores stochastic extensions to the coalescent.

birth-death phylodynamic models intrinsically rely on a stochastic model assumption. This is related to the fact that the coalescent process is derived by taking a limit to infinite population sizes where fluctuations are negligible, while our use of birth-death phylodynamic models does not involve taking such a limit. Since the relative importance of population size fluctuations can be extreme when the population size is small (even if only temporarily (Bošková, Bonhoeffer and Stadler 2014)), this means that one should take care when interpreting the results of applying coalescent models when the ancestral population is likely to have been small at some point along the tree. For example, small population sizes may occur at the start of epidemic outbreaks.

Finally, the generation times differ. Birth-death models assume exponentially distributed waiting times until birth or death events. The coalescent can be derived as a limit of the Wright-Fisher model with discrete non-overlapping generations (as demonstrated above); however, as mentioned earlier, one can also derive the coalescent under different generation time assumptions such as the Moran model (Moran 1962).

For more information about how these two models compare on exponentially growing populations, please refer to Bošková, Bonhoeffer and Stadler (2014), Volz and Frost (2014) and Stadler (2013).

9.4 Phylodynamics for non-binary trees with co-occurring internal nodes

In Chapter 6, we have considered binary tree topologies, assuming each internal node has exactly two descendants. As an extension, we mentioned trees in Section 9.1 that may have nodes with precisely one descendant, a sampled ancestor node. We did not consider trees where nodes have more than two direct descendants. Further, we only considered time trees where internal nodes all occurred at different time points: the probability of two lineages branching, or two pairs of lineages coalescing at precisely the same time, was 0 (Chapter 9 and Section 9.2).

Here, we will now mention approaches and applications where nodes may have more than two descendants or may be co-occurring in time.

There are applications where it is natural to allow parent nodes to have more than two children instead of forcing a binary tree structure. Consider, for example, the phylogeny of HIV virions. When an infected T-cell produces new virions, it does so in a "burst" that can produce tens of thousands of new viral particles simultaneously. As another example, in the case of infectious disease transmission, transmissions may occur due to "superspreading" events (imagine an infected person sneezing in a packed elevator or bus). Also, some highly fertile marine species engage in reproductive behaviours with what is known as "sweepstakes reproductive success," where there is a small but non-zero chance that a single individual may become the direct ancestor of the majority of the next generation. In all these cases, binary trees cannot capture important aspects of the population dynamics. Internal nodes with more than two descendants are called polytomies or multifurcations. Approaches have been suggested to infer trees with polytomies by pruning nodes in binary trees (Kuhn, Mooers and Thomas 2011).

While it is still uncommon, some phylodynamic methods do allow for multifurcations. One example is the multifurcating skyline plot (Hoscheit and Pybus 2019) which provides an analogue to the standard coalescent skyline plot we discussed in Section 9.2, but is based on the multifurcating generalisation to the coalescent known as the Λ-coalescent (Donnelly and Kurtz 1999; Pitman 1999; Sagitov 1999). Similarly, a method to perform tree and parameter inference under a spatially continuous generalisation of the structured coalescent known as the spatial Λ-Fleming-Viot model, which also generates polytomies, has been developed (Guindon, Guo and Welch 2016).

A related concept when considering time trees is that of “multiple mergers”. Multiple mergers are internal nodes in different parts of the tree that coincide in time. This co-occurrence may happen when we sample pathogens from within a host. At a transmission event, multiple virions may go through the transmission bottleneck founding the new infection. Initially, the strains may expand rapidly in the new host, with all the first branching events essentially co-occurring. None of the phylodynamic models described in this book thus far produce such coincidences. Models of viral phylogenies with dense sampling and transmission bottlenecks, including bottleneck-induced polytomies and multiple mergers, are in development (Stolz, Stadler and Vaughan 2024).

We note that for virions (but also for single cells), the mutations only occur at replication and not — as assumed by the available models (Chapter 5) — along branches. It remains to be investigated if this model violation causes biases, particularly under dense sampling (Section 5.1).

9.5 Accounting for population structure

So far, we assumed that the populations of interest are homogeneous, meaning that all individuals are exchangeable. In reality, most populations are structured: for instance, different individuals will have different risks of catching a particular disease based on factors such as their location, age, social group, and so on. Furthermore, based on their phenotype, different species may have a different risk for extinction.

This section will delve into the various extensions to phylodynamic models that can be used to account for underlying population structure.

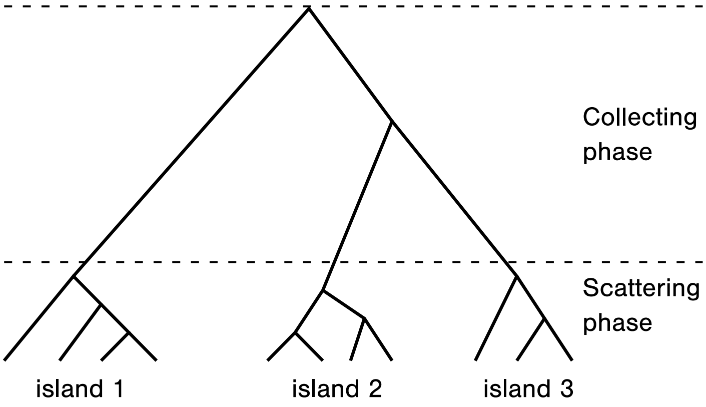


Figure 9.23: Population structure impacts the shape of the phylogeny. In this example, inspired by Pannell (2003), the ancestry of samples from three distinct island populations is shown. Weak migration between islands means that lineages from a specific island coalesce quickly with lineages ancestral to samples from the same island ("scattering phase") but slowly with lineages ancestral to samples from other islands ("collecting phase").

9.5.1 Population structure shapes phylogenies

Population structure can directly affect the shape of the resulting phylogeny. For instance, consider a toy example in which samples are collected from three distinct islands with limited migration between them. From the reconstructed phylogeny for these samples (Figure 9.23), one can see that lineages ancestral to samples from the same island coalesce rapidly, while lineages ancestral to samples from distinct islands take a lot longer to coalesce. The degree to which population structure influences the phylogeny depends heavily on the properties of the populations, so in the toy example, the migration rate plays an important role. However, it is clear that structure can affect the shape of the phylogeny to the extent that it is possible to discern this structure from the shape of the phylogeny alone. Importantly, we no longer have a uniform distribution of ranked labelled phylogenies on n tips sampled simultaneously, as we did under the unstructured birth-death and coalescent frameworks. This means that the parameters of the structured model are informed both by the branching times and the ranked tree topology (compared to only branching times under the unstructured models).

There are three important motivations for incorporating structure into phylodynamic models. Firstly, any structure unaccounted for by the model can lead to biases in the results obtained from phylodynamic analyses. For example, a basic coalescent analysis of the phylogeny shown in Figure 9.23 might conclude that the difference in coalescence rates results from a recent reduction in effective population size. In contrast, a birth-death skyline analysis might conclude that the birth rate of the population has recently increased. Incorporating structure into the

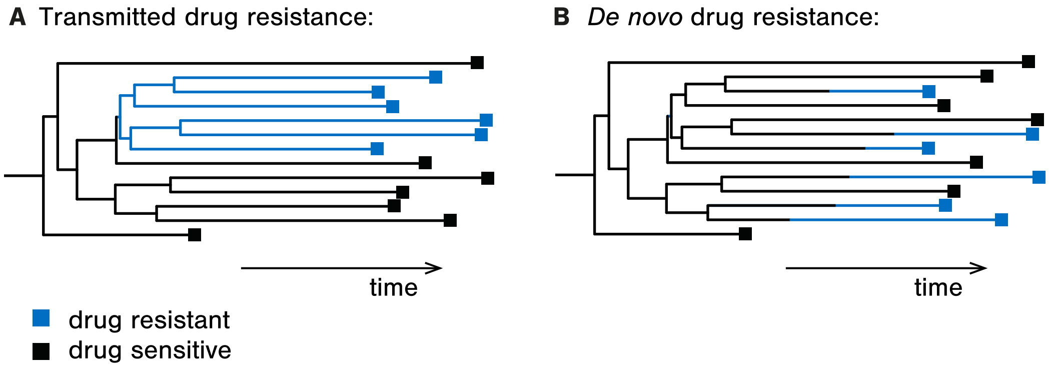


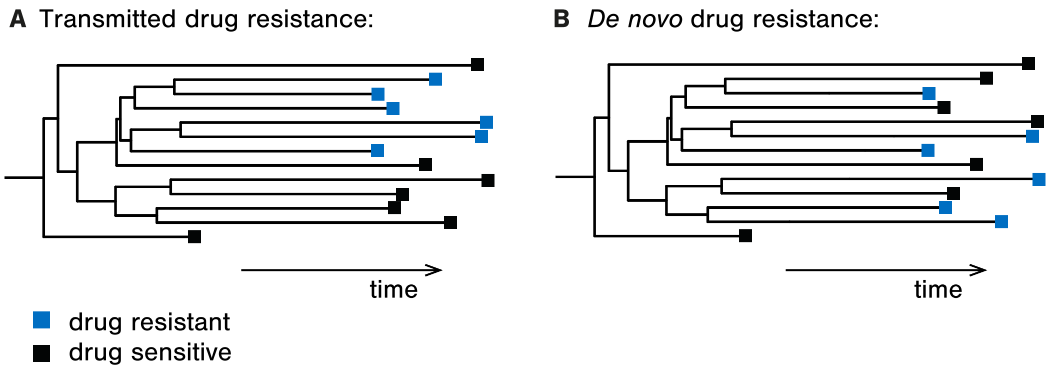
Figure 9.24: Transmission trees simulated under two different scenarios. Infected individuals are classified according to whether the infecting strain is sensitive (black) or resistant (blue) to a particular drug. **A** shows a simulated tree with drug resistance being mainly transmitted. **B** shows a simulated tree where drug resistance always evolves *de novo* within a patient.

phylodynamic models allows us to avoid this bias.

The second motivation is that incorporating structure allows us to use phylodynamic analyses to address questions relating to population structure directly. For instance, what is the migration rate between islands? What are their respective sub-population sizes? Is a particular morphological trait tied to higher rates of speciation or extinction? In the epidemiological context, one can assess whether infection rates depend on sub-populations or determine when a disease first entered a geographic region. Notably, many of these population structure questions cannot be addressed thoroughly by non-genetic time series data (data on the number of individuals through time in different population subgroups), while the genomes and their mutations and substitutions contain information about the structure through the relationships in the induced tree.

Taking the example of a pathogen with two strains, a drug-resistant strain and a drug-sensitive strain, two scenarios for the spread of the drug-resistant strain are possible: (i) transmitted drug resistance, where the drug-resistant strain is directly transmitted from patient to patient, and (ii) *de novo* drug resistance, where the drug-resistant strain is never transmitted but repeatedly arises through mutation in an already infected patient. These two scenarios and their impact on the resulting phylogenetic tree are shown in Figure 9.24.

While Figure 9.24 shows the full history, typically only the genomic sequences and the drug-resistant status of the samples in the tree are known in empirical studies. Reconstructing trees based on the genetic sequences of the samples and assigning drug resistance status (resistant or sensitive) to each tip leads to phylogenies like the ones shown in Figure 9.25. Note that this does not lead to information on the ancestral drug resistance status (branch colours as shown in Figure 9.24). Nevertheless, the reconstructed phylogenetic tree still contains valuable



A Transmitted drug resistance:

B *De novo* drug resistance:

Figure 9.25: When reconstructing transmission trees based on sequences coming from patients infected with strains that are sensitive (black) or resistant (blue) to particular drugs, then the history of drug resistance (ancestral branch colours) is missing in the reconstructed phylogenies (unlike in Figure 9.24, where we show the full information). However, blue tips forming a cluster **A** points to transmitted drug resistance while blue tips being spread across the phylogeny **B** points to frequent *de novo* evolution of drug resistance.

information about the underlying scenario: transmitted drug resistance is more likely if drug-resistant tips mostly cluster together, whereas *de novo* resistance leads to drug-resistant and drug-sensitive tips interspersed with one another (Kühnert et al. 2018b; Pečerska et al. 2021).

The third motivation is that structured models enable the quantification of selection (conceptually, this is a special case of the second motivation, but we list it separately to highlight that phylodynamics can lead to a better understanding of non-neutral processes). In the example on drug resistance, the phylogenetic trees were reconstructed based on the part of the genome not associated with drug resistance (which is, ideally, a neutrally evolving part). The drug resistance mutations are used to determine the tip label (drug-sensitive or drug-resistant). Now, the birth and death rates of the sensitive and resistant strains are estimated based on the tree and shed light on transmission fitness advantages. More generally, structured models can account for the fitness effects of any heritable phenotypic trait. Furthermore, they can quantify sequence-dependent fitness effects (Section 9.5.4.1).

9.5.2 Structured birth-death phylodynamic models

The multi-type birth-death model was developed to handle structured populations; it is an extension of the models in Section 9.1. The model contains two or more compartments or types. In epidemiology, different compartments can represent different pathogen strains, geographic locations, host risk groups, or any other pathogen or host population structure. In macro-

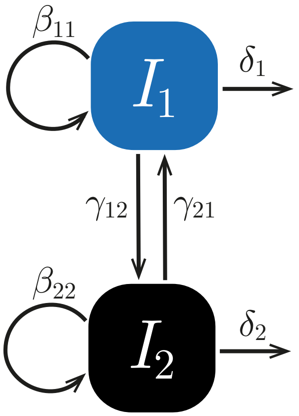


Figure 9.26: Schematic representation of a multi-type birth-death model with two compartments (states of individuals) and the dynamics associated with it (arrows labelled with rates). Here, we set both $\beta_{ij} = 0$ for $i \neq j$ and $\psi_i = 0$, thus not showing these arrows for simplicity.

evolution, different types may correspond, for example, to different species traits, different geographic locations, or different habitats.

Each compartment has its own birth rate β_{ii}, death rate δ_i, and sampling rate ψ_i. Births can also occur between compartments at rate β_{ij} for $i \neq j$, which is the rate at which individuals in i produce new individuals in compartment j. Migration rates γ_{ij} describe the rate of individuals moving from one compartment i to another j. A schematic of a model with two compartments is shown in Figure 9.26.

In the example of drug resistance, the two compartments I_1 and I_2 represent the population infected with the drug-sensitive strain and the population infected with the drug-resistant strain, respectively. β_{11} and β_{22} are the transmission rates of the drug-sensitive strain and the drug-resistant strain, and γ_{12} is the rate of resistance evolution.

In what follows, we will derive the probability density of a time tree under the multi-type birth-death model. The resulting phylodynamic likelihood is then used in parameter inference.

9.5.2.1 Deriving the probability of an individual having 0 or 1 descendants at time t

Firstly, we consider the probability $p_i(0|t)$ that no descendants of a type i individual are alive after t time units. This condition is equivalent to $p(0|t)$ in the unstructured case, with the difference being that now we also consider the type of the individual.

To compute this probability, we use the same arguments as in Section 9.1.1.1. We write the probability $p_i(0|t+\Delta t)$ in terms of $p_i(0|t)$ and use this relationship to derive a differential equation. We partition time as in Figure 9.3. In the time Δt after the start of the process, any of the following can happen:

(i) no event occurs with probability $1-(\delta_i+\sum_j(\beta_{ij}+\gamma_{ij}))\Delta t$; but since no individuals are present at $t+\Delta t$, the descendants of the original individual must go extinct within the remaining time t, which has the probability $p_i(0|t)$;

(ii) the individual dies with probability $\delta_i\Delta t$, leaving no descendants at time $t+\Delta t$;

(iii) the individual gives birth to another individual of type j with probability $\beta_{ij}\Delta t$ and both individuals must go extinct within time t, which has the probability $p_i(0|t)p_j(0|t)$;

(iv) the individual changes type (migrates) from i to j with probability $\gamma_{i,j}\Delta t$; in this case, the probability of all its descendants going extinct is the probability that all descendants of an individual of type j go extinct in the remaining time t, $p_j(0|t)$;

(v) more than one event happens, which has the probability $\mathcal{O}(\Delta t^2)$.

Combining these possibilities allows us to write down the probability of extinction:

$$\begin{aligned}p_i(0|t+\Delta t) = {} & \underbrace{\left(1-\left(\delta_i+\sum_j(\beta_{ij}+\gamma_{ij})\right)\Delta t\right)p_i(0|t)}_{(i)} \\ & +\underbrace{\delta_i\Delta t}_{ii} \\ & +\underbrace{\sum_j \beta_{ij}\Delta t p_i(0|t)p_j(0|t)}_{iii} \\ & +\underbrace{\sum_j \gamma_{ij}\Delta t p_j(0|t)}_{iv} \\ & +\underbrace{\mathcal{O}(\Delta t^2)}_{v}. \end{aligned} \tag{9.64}$$

Rearranging the terms and taking the limit $\Delta t \to 0$ yields the following differential equation:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} p_i(0|t) = &- \left(\delta_i + \sum_j (\beta_{ij} + \gamma_{ij}) \right) p_i(0|t) \\ &+ \delta_i + \sum_j \left(\beta_{ij} p_i(0|t) + \gamma_{ij} \right) p_j(0|t).\end{aligned} \tag{9.65}$$

We note that with a similar set of arguments, we can also derive the differential equation for the probability of one descendant,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} p_i(1|t) = &- \left(\delta_i + \sum_j (\beta_{ij} + \gamma_{ij}) \right) p_i(1|t) \\ &+ \sum_j \left(\beta_{ij} (p_i(0|t) p_j(1|t) + p_i(1|t) p_j(0|t)) \right) \\ &+ \sum_j \gamma_{ij} p_j(1|t).\end{aligned} \tag{9.66}$$

Unlike the differential equations for $p(0|t)$ and $p(1|t)$ in the unstructured model, the differential equations for $p_i(0|t)$ and $p_i(1|t)$ in the structured model do not have known analytical solutions. However, they can be solved using standard numerical integration techniques (Sciré et al. 2022).

9.5.2.2 Probability density of an oriented sample-typed tree

To derive the probability density of the tree under the multi-type birth-death model, we use considerations similar to those used in Section 9.1.5. As in that section, we assume complete sampling at present and that no samples have been collected before that time.

First, we define different variations of phylogenetic trees when dealing with structured models (Sciré et al. 2022). A phylogenetic tree that has type information associated with the leaves, as in Figure 9.25, is referred to as a *sample-typed tree* or *tip-typed tree*. A phylogenetic tree in which the edges at every point are annotated with the ancestral types, as in Figure 9.24 is referred to as a *branch-typed tree*. A *multi-type tree* may be either a sample-typed tree or a branch-typed tree.

In what follows, we will present a means of computing the probability density of an oriented sample-typed tree under the multi-type birth-death model. Here, we note that at a branching event with descending individuals of type i and j, the types are allocated to the left and right branches with equal probability.

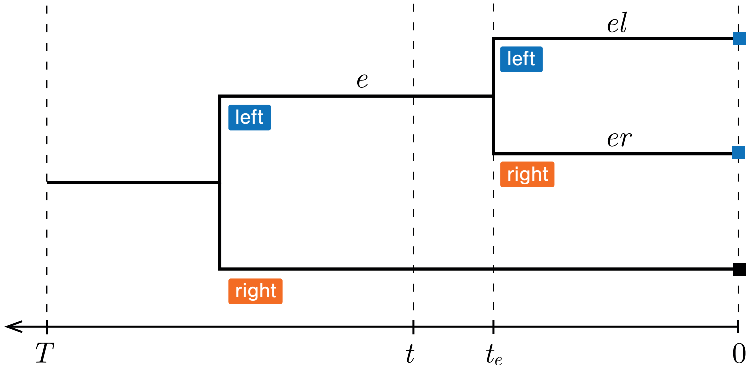


Figure 9.27: Sample-typed tree based on which we explain the probability density calculation for a tree under the multi-type birth-death model. On this tree, an arbitrary time t on edge e is marked. We consider an individual represented on edge e at time t and assume it is of type i. We denote the left and right child branches of e with el and er, respectively.

Consider an oriented sample-typed tree $\mathcal{T}^o$ and let e be a branch of $\mathcal{T}^o$. Again, let time 0 be the present, and time is reversed, meaning the origin of the tree is at time $T > 0$. Let t be some time at which branch e exists (see Figure 9.27). Let $\mathcal{T}^o_e(t)$ be the subtree descending from branch e with time of origin t. Now consider an individual of type i at time t. The probability that this individual produces $\mathcal{T}^o_e(t)$ is denoted by $g^e_i(t)$ (see Figure 9.27). By considering all events that may happen to the considered individual, we find that along branch e, $g^e_i(t)$ changes according to a differential equation of the same form as Equation (9.66):

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} g^e_i(t) = &-\left(\delta_i + \sum_j (\beta_{ij} + \gamma_{ij})\right) g^e_i(t) \\ &+ \sum_j \left(\beta_{ij} (g^e_i(t) p_j(0|t) + g^e_j(t) p_i(0|t))\right) \\ &+ \sum_j \gamma_{ij} g^e_j(t). \end{aligned} \tag{9.67}$$

At the end of branch e (time t_e, Figure 9.27), the probability density $g^e_i(t_e)$ depends on whether the node at time t_e is a leaf or a branching event:

$$g^e_i(t_e) = \begin{cases} 1 & \text{if the node is a leaf of type } i, \\ 0 & \text{if the node is a leaf of type } j \neq i, \\ \sum_j \frac{1}{2}\beta_{ij} \left(g^{el}_i(t_e) g^{er}_j(t_e) + g^{el}_j(t_e) g^{er}_i(t_e)\right) & \text{otherwise,} \end{cases} \tag{9.68}$$

where el and er are defined as the left and right child edges in the case that e gives rise to an internal node. The factor 1/2 acknowledges the equal probability of branch er starting in type i (vs. j).

The dependence of the boundary condition for internal nodes on the solutions for the branches descending these nodes allows us to numerically integrate Equation (9.67) backwards in time from each leaf, then successively combine these solutions at internal nodes until we reach the origin of the tree where the integration returns $g_i^r(T)$, where r is the branch connecting the origin and the root of tree $\mathcal{T}^o$.

The probability density $g_i^r(T)$ is the probability of observing the tree and the leaf states, given that the process began with an individual in state i at time T before the present. That is, $g_i^r(0) = P(\mathcal{T}^o|i, \boldsymbol{\beta}, \boldsymbol{\gamma}, \vec{\delta}, T)$, where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the birth and migration rate matrices, and $\vec{\delta}$ is the vector of type-specific death rates. To convert this into the probability density of the tree without conditioning on the starting state, we assume initial state probabilities π_i, where π_i is the probability of the first individual at time T in the past being in state i.

The initial state probabilities π_i need to be chosen by the user: It may make sense to fix the state of the initial individual, $\pi_j = 1$, and $\pi_i = 0$ for all states i different from j (for example, the initial individual may be assumed to be infected with a drug-sensitive strain; and not a drug-resistant strain). One can also assume $\pi_1 = \pi_2 = \ldots$ or fix the probabilities in some other way based on independent data and knowledge. Finally, if the model has a stationary distribution, one can set $\vec{\pi}$ to the stationary probabilities (Section 5.2.4). The latter was done in Maddison, Midford and Otto (2007) and Stadler and Bonhoeffer (2013).

We can then write

$$P(\mathcal{T}^o|\boldsymbol{\beta}, \boldsymbol{\gamma}, \vec{\delta}, \vec{\pi}, T) = \sum_i g_i^r(x_0)\pi_i, \tag{9.69}$$

proving the following theorem.

Theorem 9.5.1. *Consider a multi-type birth-death model for time T with birth rate matrix $\boldsymbol{\beta}$, migration rate matrix $\boldsymbol{\gamma}$, and death rate vector $\vec{\delta}$. Furthermore, consider complete extant tip sampling ($\rho = 1$) and no sampling through time ($\psi = 0$). The initial state probabilities are $\vec{\pi}$. The probability density of an oriented tree $\mathcal{T}^o$, conditioned on non-extinction ($X_T > 0$), is*

$$P(\mathcal{T}^o|T, X_T > 0, \boldsymbol{\beta}, \boldsymbol{\gamma}, \vec{\delta}, \vec{\pi}) = \sum_i \frac{g_i^r(T)\pi_i}{1 - p_i(0|T)}. \tag{9.70}$$

with $g_i^r(T)$ being evaluated according to Equation (9.67) and Equation (9.68).

Theorem 9.5.1 states the probability density of a time tree under the multi-type birth-death model with complete sampling in the present and no sampling through time. We note that the probability density of a time tree in the unstructured case (Theorem 9.1.7) was derived using a slightly different strategy. However, the same strategy as employed in the multi-type case would also prove Theorem 9.1.7. Extending this same computational approach to handle sampling through time, incomplete sampling in the present, and rates changing through time is straightforward (Sciré et al. 2022).

The expression of Theorem 9.5.1, considered as a function of the parameters, is the phylodynamic likelihood used in inference methods (see applications below).

The framework above was initially introduced for trees where all tips were sampled at the same time, aiming to model trait-dependent speciation and extinction processes leading to extant species (e.g. see Maddison, Midford and Otto (2007), FitzJohn, Maddison and Otto (2009), Goldberg, Lancaster and Ree (2011) and Goldberg and Igić (2012)). We highlight that one must carefully select which traits to use when applying this framework. To illustrate this, assume a tree evolved under some trait-dependent speciation and extinction process. Furthermore, assume that traits not influencing speciation and extinction rates are considered in a phylodynamic analysis. Then, the analysis may nevertheless estimate trait-dependent speciation and extinction rates, as the alternative (constant rates) would not explain the tree — which is shaped by a structured population — well (Rabosky and Goldberg 2015; Beaulieu and O'Meara 2016). Methods allowing for hidden traits potentially affecting speciation and extinction rates were introduced to address this problem (e.g. in Beaulieu and O'Meara (2016)).

Furthermore, models where the tip states are not known but are inferred have been presented, for example, in Stadler and Bonhoeffer (2013), Maliet, Hartig and Morlon (2019) and Barido-Sottani, Vaughan and Stadler (2020), assuming a multi-type birth-death model, and in Rabosky (2014) using a variation of the multi-type birth-death model (but see Moore et al. (2016) for limitations of the latter framework).

It is possible to use the general strategy introduced above to compute the probability density of a branch-typed tree (Sciré et al. 2022), such as the one shown in Figure 9.24. The only major difference in that case is that type-change (migration) events become additional nodes in the tree, and the differential equation for the probability distribution $g_i^e(t)$ does not allow for type change events along edge e.

9.5.2.3 Example application: geographic spread of seasonal influenza

The multi-type birth-death phylodynamic likelihood is the basis for performing phylodynamic inference on structured populations.

For example, to gain insight into the spread of the influenza virus around the globe, flu sequences — annotated according to the geographic location of the patient: northern hemisphere, southern hemisphere, or tropical area — were analysed (Kühnert et al. 2016) using a multi-type birth-death model.

The effective reproductive numbers were estimated for the three different locations over the time span of three years. The results are shown in Figure 9.28. The effective reproductive numbers for both the northern and southern hemispheres show marked seasonality: in winter, the reproductive number is above one, and in summer, it is below one. The reproductive number in the tropical area is stable, confirming that influenza is endemic in this area.

Using the phylodynamic likelihood for branch-typed trees, one can additionally infer the location of each branch in the tree. This is shown with different colours for different locations in Figure 9.29. The backbone of this tree is composed almost exclusively of tropical lineages, which indicates that the tropical area is the reservoir for the flu virus, and tropical strains start the seasonal epidemics in other locations. Viruses from the northern (or southern, respectively) hemisphere cluster together in localised epidemics, highlighting the seasonality.

9.5.3 Structured coalescent phylodynamic models

Like the coalescent distribution discussed in Section 9.2, the *structured coalescent* provides a probability distribution over sampled time trees conditional on a particular demographic (that is, population size) history. Just as in the unstructured case, the structured coalescent arises as a limiting case of a number of distinct population models, one of which is a structured extension to the Wright-Fisher model.

9.5.3.1 The structured Wright-Fisher model

The *structured Wright-Fisher model* we discuss here is described in Notohara (1990), and consists of a population of N individuals distributed among d distinct “islands”, “demes”, or “compartments”. In each deme $i = 1 \dots d$, there are a constant number of N_i individuals, with $\sum_{i=1}^{d} N_i = N$. We summarise the population sizes N_i in $\vec{N}$. The dynamics of the model occur over discrete generations, separated by a fixed generation time g. Each generation consists of two distinct phases: a migration phase and a reproduction phase. In the migration phase, individuals move freely between each pair of populations i and j with probabilities $q_{ij}g$. The migration rates are assumed to be slow enough relative to g such that $q_{ij}g < 1$ for all i, j and that $\sum_{\substack{j \\ j\neq i}} q_{ij} < 1$ for all i. In the reproduction phase, each population individually undergoes Wright-Fisher-style resampling: in deme i, each of the N_i children is assigned a parent selected uniformly at random from the N_i' members of the deme after the migration phase.

This resampling means that the population size of each deme at the end of the generation is the same as it was at the beginning. An example generation for a two-deme structured Wright-Fisher model is illustrated in Figure 9.30.

The stochastic process described above allows for the free movement of individuals between subpopulations. There is no assumption that the migration phase itself preserves the size of individual populations, even in expectation. For example, the probabilities $q_{ij}g$ and $q_{ji}g$ may be very different even in a two-deme model, meaning that under the migration process alone, the population sizes would deviate from their initial values of N_i and N_j. However, the reproduction phase of the model returns the individual population sizes to their original values. Thus, the process mimics the dynamics of a population spread across distinct demes such that

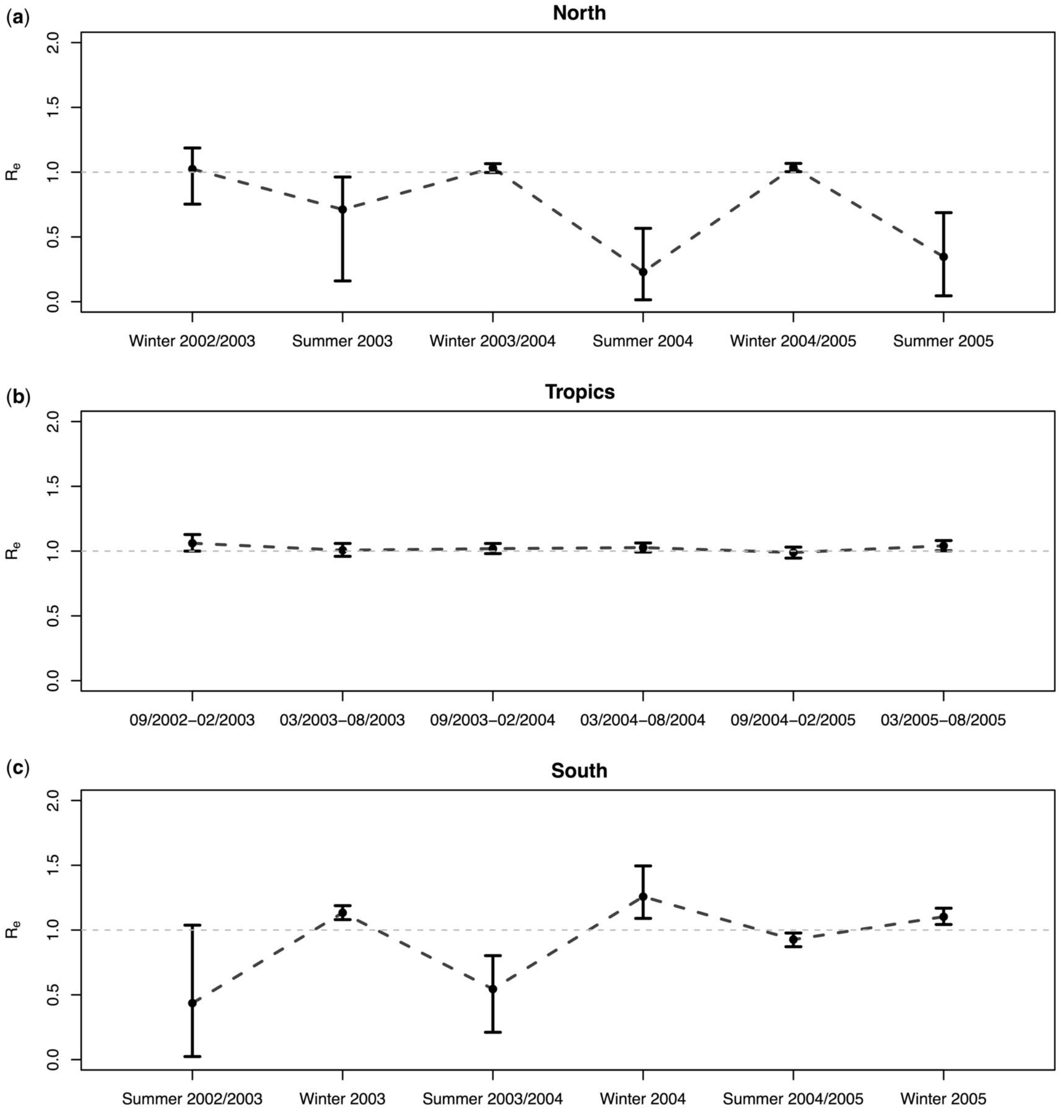


Figure 9.28: The effective reproductive number of seasonal influenza was inferred for the **(a)** northern hemisphere, **(b)** tropical areas and **(c)** southern hemisphere, under the multi-type birth-death model based on influenza sequence data. The effective reproductive number is the birth rate divided by the death and sampling rate for a particular area and season. Figure adapted from Kühnert et al. (2016).

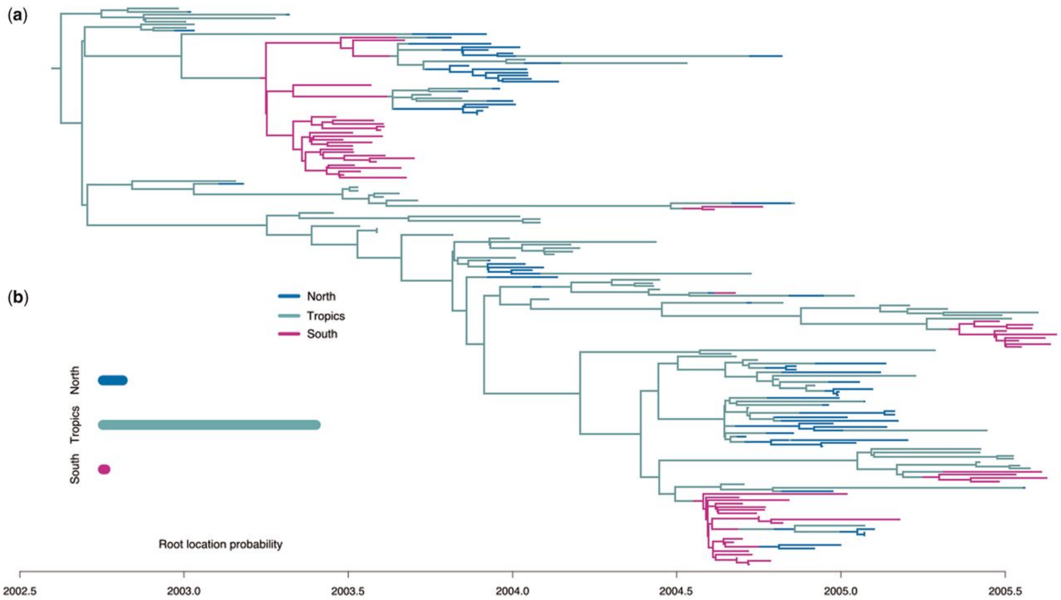


Figure 9.29: **(a)** Phylogenetic tree showing the estimated geographic spread of seasonal influenza inferred under the multi-type birth-death model. **(b)** (bottom left) shows the posterior distribution for the root location, that is, the estimated probability that the epidemic started in the northern, southern, or tropical area. Figure adapted from Kühnert et al. (2016).

the size of the population of a single deme is maintained at the deme's carrying capacity. The assumption that migration does not affect the population size distribution is equivalent to assuming that the effects of migration on the population size are only transient and that the population sizes rapidly re-equilibrate to their natural sizes (that is, carrying capacities) of $\vec{N}$.

9.5.3.2 The structured coalescent model

The structured coalescent provides a distribution for branch-typed phylogenetic trees conditional on the parameters of the model and the locations of the samples representing the tree leaves. Like the unstructured coalescent, the probability density for a branch-typed tree under the structured model can be interpreted as the probability density of a particular realisation of a backwards-in-time continuous-time Markov process. A simplified description of the derivation provided by Notohara (1990) follows.

The structured coalescent probability density of a tree is again derived from the discrete-time structured Wright-Fisher model by fixing the coalescence rate θ and taking the limit of large N (and thus small g). For this derivation, we must consider two things:

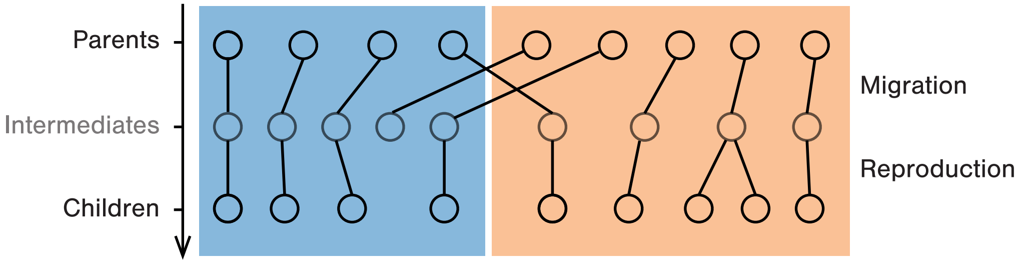


Figure 9.30: The structured Wright-Fisher model is a discrete-generation model in which each generation consists of two phases: a migration phase and a reproduction phase. This figure illustrates a possible outcome of these two phases for a two-deme model.

(i) the probability that a pair of individuals in a single deme share a common ancestor in the previous generation, and

(ii) the probability that the parent of a member of a deme in the current generation was a member of another deme in the previous generation.

In the limit of large population sizes, the first of these probabilities gives the within-deme pairwise coalescence rate, while the second gives the per-lineage backward migration rate.

The coalescence rate between a single pair. Suppose we select a pair of individuals from deme i. Under the structured Wright-Fisher model, what is the probability that these individuals share a common ancestor in the previous generation?

If we knew the total number of immigrants R_i and emigrants S_i, which respectively arrived and departed from deme i during the migration phase of the generation, we could answer this question directly using the same reasoning used to compute the probability in the unstructured case. That is, the probability that the second individual is assigned a parent identical to the first during the reproduction phase of the generation would be

$$P(\text{coalescence in } i | R_i, S_i) = \frac{1}{N_i + R_i - S_i}. \tag{9.71}$$

Thus, the probability depends on the specific movements of individuals in each generation.

At this point, it is helpful to define $\theta_i = {}^1\!/_{N_i g}$. In terms of this variable, we rewrite this probability as

$$\begin{aligned} P(\text{coalescence in } i|R_i, S_i) &= \frac{\theta_i N_i g}{N_i + R_i - S_i} \\ &= \frac{\theta_i g}{1 + (R_i - S_i)\theta_i g}. \end{aligned} \tag{9.72}$$

To determine the limiting coalescence rate of the pair of lineages (the probability of coalescence per unit of time), we must first examine the limiting behaviour of the probability distributions governing the number of immigrants and emigrants. Defining n_{ij} as the number of individuals which move from deme i to j, the assumptions of the structured Wright-Fisher model imply that n_{ij} has the following binomial distribution:

$$\begin{aligned} P(n_{ij}|\vec{N}, q, g) &= \binom{N_i}{n_{ij}} (q_{ij}g)^{n_{ij}} (1 - q_{ij}g)^{(N_i - n_{ij})} \\ &= \binom{N_i}{n_{ij}} \left(\frac{q_{ij}}{N_i\theta_i}\right)^{n_{ij}} \left(1 - \frac{q_{ij}}{N_i\theta_i}\right)^{(N_i - n_{ij})}. \end{aligned} \tag{9.73}$$

Recall that the binomial distribution approaches a Poisson distribution in the limit of a large number of trials and a small success probability (see Box 22 on page 97). This is exactly the limit we are interested in; the success probability here is $q_{i,j}/N_i\theta_i$, while the number of "trials" is N_i. Thus, for large N_i and small g (with $\theta_i = 1/(N_i g)$) we find that the probability distribution for n_{ij} approaches a Poisson distribution with mean q_{ij}/θ_i. That is,

$$P(n_{ij}|\vec{N}, q, \theta_i) \simeq e^{-q_{ij}/\theta_i} \frac{(q_{ij}/\theta_i)^{n_{ij}}}{n_{ij}!}. \tag{9.74}$$

The total number of immigrants into deme i is given by

$$R_i = \sum_{\substack{j \\ j \neq i}} n_{ji}. \tag{9.75}$$

Since the sum of Poisson-distributed random variables is also Poisson-distributed (Box 21 on page 96), R_i is Poisson-distributed with a mean (and variance) equal to $\sum_{\substack{j \\ j \neq i}} q_{ji}/\theta_j$. Similarly, the number of emigrants out of i is given by $S_i = \sum_{\substack{j \\ j \neq i}} n_{ij}$ and is therefore also Poisson-distributed with a mean (and variance) given by $\sum_{\substack{j \\ j \neq i}} q_{ij}/\theta_i$.

Since the mean and variance of R_i and S_i depend only on $\vec{\theta}$ and q, they remain fixed in the coalescent limit. Thus, the coalescence rate between the pair of individuals converges in probability to

$$\begin{aligned}\lim_{g\to 0}\frac{P(\text{coalescence in } i)}{g} &= \lim_{g\to 0}\frac{\sum_{R_i,S_i} P(\text{coalescence in } i|R_i,S_i)P(R_i,S_i)}{g}\\ &\overset{(9.72)}{=} \theta_i \lim_{g\to 0}\sum_{R_i,S_i}\frac{P(R_i,S_i)}{1+(R_i-S_i)\theta_i g}\\ &= \theta_i = {}^{1}\!/_{(N_i g)}. \end{aligned} \tag{9.76}$$

Since the reproduction phase follows the migration phase, members of different populations cannot share a parent in the previous generation. Hence, the coalescence rate between pairs of individuals in distinct demes is zero.

Thus, in general, the rate of coalescence between a pair of individuals in demes i and j is given by θ_i if $i = j$ and is 0 otherwise.

The backwards migration rate of an individual. Now, suppose we select a single individual from deme i. What is the probability that this individual has a parent who was a deme $j \neq i$ member in the previous generation?

To do this, we introduce the variables Z_p and Z_c to represent the demes of the parent and the child, respectively, and derive the probability $P(Z_p = j|Z_c = i)$. The expression for $P(Z_p = j|Z_c = i)$ can be rearranged using the rules of conditional probability (see Section 1.3.1) in the following way:

$$\begin{aligned}P(Z_p = j|Z_c = i) &= \frac{P(Z_c = i|Z_p = j)P(Z_p = j)}{P(Z_c = i)}\\ &= \frac{P(Z_c = i|Z_p = j)\times P(Z_p = j)}{\sum_k P(Z_c = i|Z_p = k)\times P(Z_p = k)}.\end{aligned} \tag{9.77}$$

The terms on the right-hand side of this equation are directly provided by the model, so we can write

$$\begin{aligned}P(Z_p = j|Z_c = i) &= \frac{q_{ji}g^{N_j}/N}{\left(\sum_{\substack{k\\k\neq i}} q_{ki}g^{N_k}/N\right)+\left((1-\sum_{\substack{k\\k\neq i}} q_{ik}g)^{N_i}/N\right)}\\ &= \frac{q_{ji}\theta_j^{-1}}{\left(\sum_{\substack{k\\k\neq i}} q_{ki}\theta_k^{-1}\right)+(N_i-\sum_{\substack{k\\k\neq i}} q_{ik}\theta_i^{-1})}.\end{aligned} \tag{9.78}$$

Note that the expression in the right bracket of the denominator refers to the case $k = i$. The backward-time rate that a lineage in deme i migrates to deme j is then given by

$$\begin{aligned} M_{ij} = \lim_{g\to 0} \frac{P(Z_p = j | Z_c = i)}{g} &= \lim_{g\to 0} \frac{q_{ji}\theta_j^{-1}}{\sum_{\substack{k\\k\neq i}} q_{ki}\theta_k^{-1} g + (N_i g - \sum_{\substack{k\\k\neq i}} q_{ik}\theta_i^{-1} g)} \\ &= \lim_{g\to 0} \frac{q_{ji}\theta_j^{-1}}{\sum_{\substack{k\\k\neq i}} q_{ki}\theta_k^{-1} g + (\theta_i^{-1} - \sum_{\substack{k\\k\neq i}} q_{ik}\theta_i^{-1} g)} \\ &= \frac{q_{ji}\theta_j^{-1}}{\theta_i^{-1}} \\ &= \frac{q_{ji}N_j}{N_i}. \end{aligned} \tag{9.79}$$

9.5.3.3 The structured coalescent process and the probability of a branch-typed tree

In the previous section, we derived the rate of coalescence of two individuals in the same deme i to be $\theta_i = {}^1/_{N_i g}$, the rate of coalescence of two individuals in distinct demes to be 0, and the backwards rate of migration from deme i to deme j ($i \neq j$) to be $M_{i,j} = \frac{q_{ji}N_j}{N_i}$. We can now formulate the structured coalescent as a backwards-time stochastic process. To do this, we introduce the vector $\vec{k}$ whose elements k_i represent the number of sampled lineages belonging to deme i.

We generalise the pairwise coalescence and backwards-time migration rates to provide the coalescence and backwards migration rates for arbitrary numbers of lineages.

Suppose we have k_i sampled lineages in deme i. As the coalescent limit only allows for independent pairwise coalescences (coalescence between three or more lineages occurs with probability 0 in this limit), we can multiply the pairwise coalescence rate $\theta_i = {}^1/_{N_i g}$, by the number of pairs of sampled lineages in this deme, to get the following expression for the total coalescence rate in this deme:

$$\binom{k_i}{2}\frac{1}{N_i g}. \tag{9.80}$$

Similarly, we can obtain the total rate of backwards migration of sampled lineages from deme i to some other deme j by multiplying the per-lineage backwards migration rate in that direction, $M_{ij} = {}^{(q_{ji}N_j)}/_{N_i}$ by the number of sampled lineages in i to get

$$k_i M_{ij}. \tag{9.81}$$

With these transition rates, we can write down the probability density for a branch-typed tree T under the structured coalescent. Consider the branch-typed tree shown in Figure 9.31.

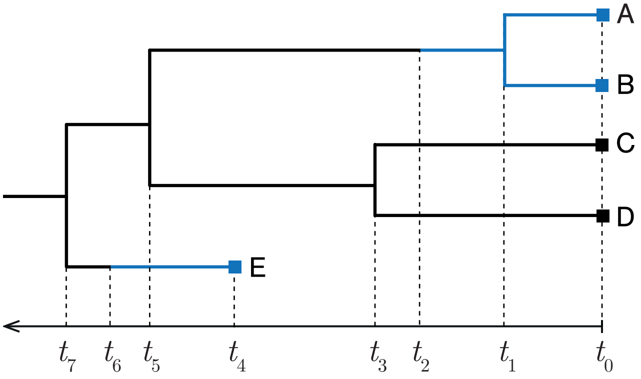


Figure 9.31: A branch-typed tree for which the tree probability density calculation under the structured coalescent is explained. The tree has the event times $t_0, \ldots, t_7$ (coalescence, migration and sample times) marked. The time variable increases into the past.

The probability of this tree under a two-deme structured coalescent model conditional on the sample types C and the model parameters can be written as follows:

$$\begin{aligned} f(\mathcal{T}|\vec{N}g, \mathbf{M}, C) = \exp&\left(-(t_1 - t_0)\left(\binom{2}{2}\frac{1}{N_1 g} + \binom{2}{2}\frac{1}{N_2 g} + 2M_{12} + 2M_{21}\right)\right)\frac{1}{N_1 g} \\ \times \exp&\left(-(t_2 - t_1)\left(\binom{1}{2}\frac{1}{N_1 g} + \binom{2}{2}\frac{1}{N_2 g} + M_{12} + 2M_{21}\right)\right) M_{12} \\ \times \exp&\left(-(t_3 - t_2)\left(\binom{3}{2}\frac{1}{N_2 g} + 3M_{21}\right)\right)\frac{1}{N_2 g} \\ \times \exp&\left(-(t_4 - t_3)\left(\binom{2}{2}\frac{1}{N_2 g} + 2M_{21}\right)\right) \\ \times \exp&\left(-(t_5 - t_4)\left(\binom{1}{2}\frac{1}{N_1 g} + \binom{2}{2}\frac{1}{N_2 g} + M_{12} + 2M_{21}\right)\right)\frac{1}{N_2 g} \\ \times \exp&\left(-(t_6 - t_5)\left(\binom{1}{2}\frac{1}{N_1 g} + \binom{1}{2}\frac{1}{N_2 g} + M_{12} + M_{21}\right)\right) M_{12} \\ \times \exp&\left(-(t_6 - t_5)\left(\binom{2}{2}\frac{1}{N_2 g} + 2M_{21}\right)\right)\frac{1}{N_2 g}. \end{aligned} \tag{9.82}$$

Each line of the right-hand expression corresponds to the probability of seeing no event in the associated time interval, followed by the probability (density) of the observed event that terminates the interval. Just as in the unstructured case presented in Section 9.2.2.3, the absence of the binomial coefficients in the event probabilities is due to these terms representing the probability that a particular pair coalesces or a particular lineage migrates. Also, as in the unstructured case, the tree probability is conditional on the sample times and locations; thus, the sampling events do not contribute directly to the probability but instead alter the number of lineages at particular times. Note that the above equation uses the generalised binomial coefficient definition from Box 2 on page 25, in which $\binom{a}{b} = 0$ when $b > a$.

In general, the probability density of a branch-typed tree is

$$f(\mathcal{T}|\vec{N}g, \mathbf{M}, C) = \prod_{l=1}^{L} \exp\left(-(t_l - t_{l-1}) \sum_{i=1}^{d} \left(\binom{k_i^l}{2} \frac{1}{N_i g} + \sum_{j=1}^{d} k_i^l M_{ij}\right)\right) \times \prod_{i=1}^{d} \left(\left(\frac{1}{N_i g}\right)^{\nu_i^c} \prod_{j=1}^{d} (M_{ij})^{\nu_{ij}^m}\right), \quad (9.83)$$

where $l \in [0, L]$ indexes the L intervals between unique coalescence, migration and sample times in order of increasing age, t_l is the time at the oldest end of the interval l, and we define $t_0 = 0$ as the age of the most recent leaf. We further define k_i^l as the number of lineages in deme i in the interval between t_l and t_{l-1}, ν_i^c as the number of coalescence events between lineages in deme i, and ν_{ij}^m as the number of migration events on the tree from deme i to deme j backwards in time.

If the tree and the ancestral migration history were perfectly known, this probability density could be used on its own as the basis for a maximum likelihood inference scheme to infer migration rates M_{ij} and effective population sizes N_i.

In practice, however, the ancestral migration history is rarely known. Thus, the structured coalescent is often used as one component of a larger inference scheme where the tree and the ancestral migration history are inferred together with the migration rates and effective population sizes (see Chapter 10 for details).

9.5.3.4 Expected coalescence times under 2-deme symmetric model

Unfortunately, the complexity of the structured coalescent model means that few results can be proved analytically. However, among those that can be proven is a particularly elegant result regarding the expected time to the most recent common ancestor (tMRCA) of two samples in a structure coalescent model with two demes, equal sub-population sizes $N_1 g = N_2 g = Ng$ and symmetric backwards migration rates $M_{12} = M_{21} = m$.

To derive this result (which can also be found, for example, in Hein, Schierup and Wiuf (2005)), we define τ_S as the expected tMRCA for two samples drawn from the same population and τ_D as the expected tMRCA for a pair of samples drawn from distinct populations (that is, one sample from each deme).

To derive expressions for these quantities, we begin by relating them algebraically. The relationships we write down rely heavily on the properties of exponential distributions and

Poisson processes laid out in Box 21 on page 96. In particular, we use the fact that the mean value of an exponentially distributed time t with rate parameter λ is given by $1/\lambda$, as well as the fact that the minimum of two event times t_1 and t_2 which are individually exponentially distributed with rates λ_1 and λ_2 is itself exponentially distributed with rate $\lambda_1 + \lambda_2$.

Firstly, consider τ_D, the expected time until coalescence for lineages ancestral to samples in different demes. Since lineages in different demes cannot directly coalesce, a migration in either of the two possible directions has to occur first. Thus, τ_D must include the expected waiting time until one of the two possible migrations occurs. After such a migration, the two lineages are in the same deme, so the remaining time to coalescence is τ_S. That is,

$$\tau_D = \frac{1}{2m} + \tau_S, \tag{9.84}$$

where $1/2m$ is the expected time for either migration to occur since this time is exponentially distributed with rate $m + m$ (that is, one m for each migration direction).

We can similarly decompose τ_S into the expected time until any event occurs and then deal with the two possibilities (migration or coalescence) individually. This gives

$$\tau_S = \frac{1}{2m + 1/Ng} + p_M \tau_D + p_C \cdot 0, \tag{9.85}$$

where p_M and p_C are the respective probabilities of the event being a migration or a coalescence, and the factor following the migration probability is τ_D, as again, the two lineages are in distinct demes. The factor following p_C is zero because there is no additional waiting time as the coalescence has already occurred. Substituting in $p_M = \frac{2m}{(2m+1/Ng)}$ (which is the ratio of the total migration rate to the sum of all of the rates) and eliminating the term with the zero factor yields:

$$\tau_S = \frac{1}{2m + 1/Ng} + \frac{2m\tau_D}{2m + 1/Ng}. \tag{9.86}$$

Solving the recursion relation defined by Equations (9.84) and (9.86) then gives the following values for the mean tMRCAs:

$$\tau_S = 2Ng, \tag{9.87}$$

$$\tau_D = \frac{1}{2m} + 2Ng. \tag{9.88}$$

As expected, τ_D increases without bound in the limit of low migration since there must be at least one migration before coalescence can occur under the structured coalescent model. Interestingly, however, τ_S, the expected time for a pair of lineages from the same deme to coalesce in this symmetric two-deme model, is independent of the migration rate m. Moreover, this is exactly the same expectation as in an unstructured model with a total population size of $2N$. However, this correspondence only holds in the mean, as the variance in the tMRCA, even in this simple model, depends on the migration rate and is thus distinct from the unstructured model.

9.5.3.5 Approximations to the structured coalescent

While the probability of a branch-typed tree can easily be written down, computed, and used directly for inference (as in the following example), using it in practice is challenging since this likelihood requires explicit consideration of ancestral types — information that is rarely directly observable. This means that, in practice, we usually need to sum over all the possible ancestral states. Since there is no general analytical solution to this problem, this summation has to be computed numerically, dramatically increasing the computational complexity of the problem.

Several authors have proposed approximations to the structured coalescent to combat this complexity, allowing this summation to be performed much more efficiently, at the cost of introducing slight deviations from the true structured coalescent model. All involve attempts at replacing the ancestral types in a branch-typed tree with the probabilities that lineages occupy particular states given the observed types at the leaves. The approximations of Volz (2012) and de Maio et al. (2015) (the latter method is implemented as a package BASTA (https://bitbucket.org/nicofmay/basta-bayesian-structured-coalescent-approximation) in BEAST2 (https://www.beast2.org/)) both compute lineages type probabilities under the assumption that the migration of one lineage is completely independent of all other lineages. An improved approximation of Müller, Rasmussen and Stadler (2017) (implemented in the software package MASCOT (https://github.com/nicfel/The-Structured-Coalescent), also as a part of BEAST2) acknowledges that the migration of one lineage is influenced by the location of other lineages and their coalescent probabilities.

These methods are extremely influential in the practical application of structured coalescent models.

9.5.3.6 Example application: geographic spread of seasonal influenza

As in Section 9.5.2.3, we demonstrate the use of the structured coalescent by again considering the case of the worldwide circulation of the influenza virus. Vaughan et al. (2014) analysed a similar dataset to that used in Section 9.5.2.3. They used a three-deme structured coalescent model, with the three effective population sizes accounting for the differing infectious pool sizes in the three sample locations: New Zealand (southern hemisphere), Hong Kong (tropics), and New York (northern hemisphere). The results, shown in Figure 9.32, agree with the results of the multi-type birth-death model in Figure 9.29, suggesting that the "trunk" of the transmission tree is predominantly in the tropics. In addition, results from the coalescent model analysis provide a direct inference of Ng for the three locations. The ordering of these

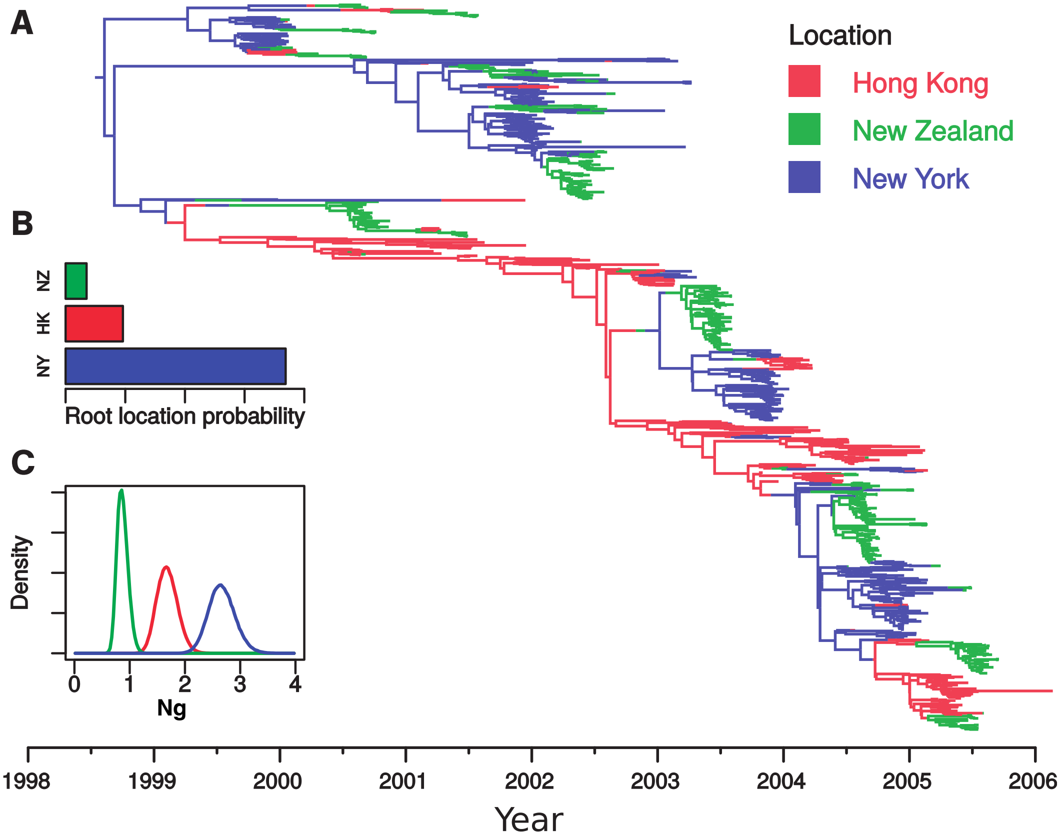


Figure 9.32: Structured coalescent model analysis of influenza sequences. A three-deme structured coalescent model is assumed to infer **A** ancestral locations, **B** the root location, **C** effective population sizes, and migration rates (not pictured), using influenza genetic sequence data collected at different times from three locations – tropics: Hong Kong, southern hemisphere: New Zealand, northern hemisphere: New York. Figure adapted from Vaughan et al. (2014).

Ng values reflects the ordering of human population sizes, with the New Zealand deme having the smallest Ng and New York having the highest. This ordering of Ng values implies a similar ordering of effective population sizes, assuming the generation time g is comparable between these locations.

9.5.4 Related structured models

9.5.4.1 Neutral trait evolution (phylogeography)

So far, in this section, we have focused on models for structured populations that allow different compartments or populations to have different demographic characteristics. For instance, members of the drug-resistant and drug-sensitive compartments in the model shown in Figure 9.26 are assumed to have different removal or recovery rates. Similarly, members of different demes in a structured coalescent model coalesce at different rates due to differences in population sizes.

A large class of structured models focuses on the special case where the deme/compartment does not affect quantities such as birth, death, and coalescence rates. Such models are sometimes referred to as "neutral trait" models, where we identify compartment membership with a particular trait and claim that the value of this trait does not affect reproductive success.

Such models are popularly applied to problems in *phylogeography*: the study of the relationship between phylogeny and geography. A significant contribution was made by Lemey et al. (2009) providing a widely-used method under such a model.

The idea behind such "neutral trait" models is fairly simple: a tree is assumed as given (meaning we can assume it was generated, for example, by any of the unstructured birth-death or coalescent models for the tree generation process). Traits are assumed to evolve forward in time along the branches of the tree according to a continuous-time Markov chain with fixed transition rates. This is perfectly analogous to how we model nucleotide substitutions on a phylogeny. In fact, such models are mathematically equivalent to appending a single additional site to the genetic MSA that evolves according to a special substitution model. As such, computing the probability of the trait distribution among the samples can be done directly using Felsenstein's pruning algorithm and thus is highly efficient.

Importantly, the neutrality assumption also extends to the sampling process: such models generally assume that the trait value carried by an individual has no bearing on its probability of being sampled. Violating this assumption is known to result in biased estimates of phylodynamic parameters (de Maio et al. 2015).

9.5.4.2 Models with continuous structure

The structured population models we have described so far have all had one thing in common: they assume that a population may be divided into discrete sub-populations, within which individuals are identical. However, many populations are better described by models with a continuum of compartments. An obvious example is the spatial distribution of a population. Although we can approximate this distribution using discrete demes, as in the structured Wright-Fisher model, it is generally more natural to assign a unique location to every individual.

A variety of phylodynamic models allow for continuously structured populations. One of the simplest is the continuous version of the neutral trait model discussed in Section 9.5.4.1 (Lemey et al. 2010). These models, commonly referred to as *continuous phylogeography* models, are neutral in the sense that the continuous structure does not affect the birth and death rates. An individual has the same rates no matter which continuous structure value characterises it.

Continuous phylogeography models have been widely applied in an effort to understand the geographical spread of populations of all kinds. For example, in the traditional context of species phylogeography, this approach has been employed in the study of the spread of plant (Ronikier et al. 2023) and animal (Malleret et al. 2022) populations over long timescales. Continuous phylogeographic analyses have also been used to infer fine-grained details of the spatial dynamics of diverse pathogens such as yellow fever virus in Brazil (Faria et al. 2018), and the SARS-CoV-2 in the United Kingdom (Kraemer et al. 2021), as well as the global spread of avian influenza virus (The Global Consortium for H5N8 and Related Influenza Viruses 2016), to name just a few. Finally, such models have also been used to infer properties of the geographical dynamics of languages, including identifying the origin of the Indo-European family (Bouckaert et al. 2012) and understanding the expansion of the Bantu linguistic family out of West Central Africa (Grollemund et al. 2015).

The following two approaches do not make the neutrality assumption; the trait may affect the reproduction process. Maliet, Hartig and Morlon (2019) introduced a model where, upon birth, the birth rate may shift following the law of some continuous distribution. This model thus makes it possible to have a continuum of lineage-specific birth rates. A continuous variant of the structured coalescent model also exists, known as the spatial Λ-Fleming-Viot process (Barton, Etheridge and Véber 2010). A phylodynamic inference scheme based on this model was developed by Guindon, Guo and Welch (2016), although its computational demands have limited its use (Guindon and de Maio 2021).

9.5.4.3 Genotypes affecting reproductive fitness

One way in which many real populations exhibit population structure is via the effects of natural selection. Any given biological population may have diversity in the fitness of its members, leading to a heterogeneity of birth, death, and coalescence rates in the population. Although this is often ignored for practical reasons, this can result in biases in inference results (Neher and Hallatschek 2013).

One direct approach to incorporating this fitness variation is treating fitness classes as compartments in the multi-type birth-death model introduced above. While this solves the problem in principle, the large number of compartments necessary to represent the possible fitness variation in a real population requires sophisticated approximations, such as those presented in the work of Rasmussen and Stadler (2019), to make phylodynamic inference under such models feasible.

9.5.4.4 Combating over-parameterisation in multi-type models

Analyses using multi-type phylodynamic models of all forms suffer from the common problem that the number of unknown parameters that need to be inferred increases rapidly with the number of distinct types or demes considered. Consider, for instance, a multi-type birth-death model parameterised by type-specific birth and death rates and migration rates. In the most general case, the number of unique parameters increases quadratically with the number of types. Considering that an increase in available data does not necessarily mirror this increase in model complexity, analysing samples drawn from populations involving more than a handful of types can easily present an insurmountable challenge.

One approach to addressing this problem is to employ so-called "regularisation" approaches, which, for example, may apply a penalty to non-zero rate parameters to ensure the total number of rates that need to be estimated is kept as low as possible while still having sufficient rates to explain the data. This approach is used by the original discrete neutral trait method of Lemey et al. (2009) to keep the number of type-transition rate parameters low, and the regularisation allows this method to be used with many distinct types.

Furthermore, one can treat the different birth and death rates not as independent parameters but rather as discrete categories from a continuous distribution, similar to how rate variations between sites can be handled in a substitution model (see Section 5.5). In this situation, the parameters are the number of types d and the mean and variance of the chosen distribution. Based on these three parameters, the continuous distribution is discretised in d parts, leading to a rate for each type. Thus, the number of parameters no longer depends on the number of types. This approach was implemented in Höhna et al. (2019).

Another approach to dealing with the explosion in model complexity is to treat the model parameters as possible covariates of external environmental factors. For instance, assuming migration rates between spatial compartments are inversely proportional to the geographic distance between the compartment centroids. This approach, which also extends to skyline plot models with time-dependent rates, not only addresses the practical problems of over-parameterisation but also allows us to investigate possible explanations for rate variation directly. A popular example of this approach is the analysis of global patterns of seasonal influenza transmission conducted by Lemey et al. (2014), where migration rates of a neutral trait evolution model are tied probabilistically to several possible explanatory variables, including distance between locations, population density, and frequency of air travel between locations. Another example is given by Müller, Dudas and Stadler (2019), who used the structured coalescent to model the phylodynamics of the Ebola virus in Sierra Leone, linking effective population size dynamics to observed case counts and migration rates to geographic distances between districts.

9.5.4.5 Ecological models

Several multi-compartment phylodynamic models (Etienne et al. 2011; Volz et al. 2009; Rasmussen, Ratmann and Koelle 2011; Leventhal et al. 2014) were developed to assess ecological and epidemiological questions. In ecology, a carrying capacity was introduced as a separate compartment. One compartment comprises the species that are also tracked in the phylogeny, and the second comprises the free niches that affect the dynamics of the first compartment. Upon extinction or speciation of a species, the number of free niches increases or, respectively, decreases by one. The speciation rate depends on the number of free niches. In epidemiology, the considered models are so-called SIR-type models where infected individuals (compartment I) are tracked in a phylogeny, while the number of susceptible individuals (compartment S) may affect dynamics through density-dependent effects.

Importantly, in this book, we track single populations in a phylogeny, such as species or infected individuals. Potentially, the multi-compartment models could be generalised to multi-type Lotka-Volterra (Lotka 1910; Volterra 1928) models, fully embedding population dynamic approaches within phylodynamics by appreciating the interactions and competitions between different populations. In particular, under Lotka-Volterra models, several populations that span separate trees (e.g. a predator and a prey tree) would be tracked. Phylodynamic models accounting for these different trees can shed light on the interaction dynamics between these populations.

10 Bayesian inference

In the previous chapters, we discussed the key components that go into a phylogenetic and phylodynamic analysis. First, we have discussed how sequence data are obtained and processed. Then, we discussed the models that give rise to sampled sequence data: models for the mutation and substitution processes and models for the generation of trees by populations of reproducing individuals. Finally, we have discussed how these models, coupled with the data, can be used in a statistical framework to infer phylogenetic relationships and phylodynamic model parameters. In particular, we highlighted inference methods using algorithmic (such as UPGMA) or optimisation (such as maximum likelihood) approaches.

These presented methods of inference have some limitations. Firstly, assessing uncertainty is challenging, often requiring external methods such as bootstrap resampling. In particular, a phylodynamic analysis is done assuming a fixed phylogenetic tree, that is, assuming no uncertainty in the tree. Secondly, these methods provide no mechanism to incorporate prior information into an analysis other than fixing parameters to known values. Finally, performing model selection is not straightforward. For example, when models are not nested, the thresholds for model selection criteria such as AIC (Section 7.3) can be difficult to interpret. On the other hand, when nested models are compared in a pairwise fashion, correcting for multiple testing can be challenging.

This chapter introduces the Bayesian phylogenetic and phylodynamic inference framework as an alternative. Here, the phylogenetic tree with the substitution model and phylodynamic parameters are estimated simultaneously based on the genetic sequence data. With this approach, an inference result is always coupled with an estimate of the uncertainty, and the inclusion of prior information is a natural and important part of the analysis. Further, explicit model selection can often be reduced to comparing easily understandable model probabilities or implicitly averaging over possible model choices.

This chapter first introduces the fundamentals of Bayesian inference and then shows how genetic sequence data can be analysed within this framework, assuming the data have a shared phylogenetic history.

10.1 Bayesian theory

Suppose we have data D from which we want to learn something about a parameter θ under some model M. The model is generative and stochastic, meaning it connects the parameter and the data probabilistically via $P(D|\theta, M)$. How can we proceed?

We have seen several approaches to this general problem in previous chapters of this book. We might, for instance, use some estimator $\hat{\theta}_f = f(D)$ with well-known statistical properties. We can also use analytical or computational means to produce confidence intervals containing the truth in a certain chosen percentage of all cases (often 95%). Estimators and confidence intervals are often derived using the likelihood function $P(D|\theta, M)$ (e.g. maximum likelihood estimators, see Box 25 on page 116) but not necessarily (e.g. the UPGMA tree reconstruction algorithm, see Section 6.3.1.1).

The Bayesian approach to this general problem is as follows: suppose that what we want to know about θ is best expressed as a probability distribution of θ given the model and our data, $P(\theta|D, M)$. Knowing this distribution would allow us to compute probabilities for any characteristic of θ directly.

We proceed to use the rules of probability described in Section 1.3 to express $P(\theta|D, M)$ in terms of the likelihood function:

$$P(\theta|D, M) = \frac{P(\theta, D|M)}{P(D|M)} = \frac{P(D|\theta, M)P(\theta|M)}{P(D|M)}. \qquad (10.1)$$

This is known as the *Bayes' rule*[1], and it establishes a link between what is known about θ in the absence of D, $P(\theta|M)$, and what is known after the data have been taken into account, $P(\theta|D, M)$.

The different components of Bayes' rule have specific names:

(i) $P(\theta|D, M)$ is the *posterior probability distribution* or the *posterior* of the parameter θ,

(ii) $P(D|\theta, M)$ is the *likelihood* for θ under the model M,

(iii) $P(\theta|M)$ is the *prior probability distribution*, or simply the *prior* for θ, and

(iv) $P(D|M)$ is the *marginal likelihood* of the model M, also known as the *evidence* for that model.

It is important to note that these names identify the roles these distributions play in a particular analysis rather than any intrinsic property of the distributions. That is, while $P(\theta|D, M)$ plays the role of the posterior in the case above, it can also be used as the prior for a subsequent analysis, as we will discuss later.

Note that the model M constitutes not only the assumptions underlying the likelihood function $P(D|\theta, M)$ but also the prior knowledge of the model parameters θ embodied in $P(\theta|M)$.

[1] Bayes' rule is named for Reverend Thomas Bayes, a Presbyterian minister who derived and applied the rule in an essay read to the Royal Society in 1763 (Jaynes 2003).

The marginal likelihood is so named because it can be considered a result of integrating out (marginalising) the model parameters θ in the likelihood function:

$$P(D|M) = \int_{\theta} P(D|\theta, M)P(\theta|M)\,\mathrm{d}\theta. \tag{10.2}$$

Thus, a practical interpretation of the role of the marginal likelihood in Equation (10.1) is that of a normalising constant in an analysis (as D and M are constants).

The marginal likelihood is often difficult to compute, especially when θ is a vector with many dimensions (the model contains many parameters), and the integral cannot be solved analytically. Computing the marginal likelihood is the central challenge of many Bayesian inference problems. Later in this chapter, we will discuss numerical techniques that allow us to proceed without directly computing this quantity.

10.1.1 Maximum likelihood vs. Bayesian inference

Due to the simplicity of the derivation above, it is easy to assume that the difference between Bayesian inference and non-Bayesian inference is minimal. After all, the Bayes' rule is a simple consequence of conditional probabilities.

It is important to note that the principal distinction between the inference strategies discussed in this chapter and those presented in earlier chapters has to do with the interpretation of probability. In previous chapters, our use of probability has been restricted to situations where random variables are used to represent the outcome of some random process, $P(D|\theta, M)$: dice rolls, sequence evolution, and so on. In other words, the probability of an outcome can be regarded as the relative frequency with which that outcome occurs. The available data are an outcome of the process.

In contrast, by writing $P(\theta|D, M)$, we are using a random variable to describe a parameter of the model. In our context, it is the parameter of a model giving rise to our data. In doing so, we represent a distribution of a state of knowledge rather than the physical properties of a system. The question immediately arises: "Whose state of knowledge?" The answer to this question is simply, "Anyone with access to the same information about the parameter."

A consequence of the Bayesian framework is that it is completely permissible for different analyses to give different results if they are subject to different information, as encoded in the prior $P(\theta|M)$.

Be aware that in cases where a single model is assumed during the analysis, it is common to avoid explicitly including M in the formulation of Bayes' rule. However, one must take care when omitting this term as its absence can lead to confusion (e.g. the term $P(D|M)$ in the denominator is arguably clearer than $P(D)$ on its own).

10.1.2 Example: Bayesian inference of pairwise genetic distance

In Section 5.4, we discussed estimating the genetic distance, measured in the expected number of substitutions, between a pair of sequences under the JC69 substitution model. We used a method of moments estimator (Section 5.4.2) as well as a maximum likelihood estimator (Section 5.4.3). We will now discuss a Bayesian approach to the same problem.

Suppose we have a pair of aligned sequences of length L, and suppose the sequences differ at S sites. Under the JC69 model, the probability of observing a difference at a single site after time t assuming an evolutionary rate λ is given by Equation (5.56),

$$p(d) = \frac{3}{4} - \frac{3}{4}e^{-\frac{4}{3}d}, \tag{10.3}$$

where d is the genetic distance between the sequences, $d = 3\lambda t$.

Since each site evolves independently under this model, we can write the likelihood function:

$$P(S|d, M_L) = \binom{L}{S} p(d)^S (1-p(d))^{L-S} = \binom{L}{S}\left(\frac{3}{4} - \frac{3}{4}e^{-\frac{4}{3}d}\right)^S \left(\frac{1}{4} + \frac{3}{4}e^{-\frac{4}{3}d}\right)^{L-S}. \tag{10.4}$$

Here, M_L represents our model, the JC69 model on L sites, and our prior assumption on the model parameter d. In Section 5.4.3, the likelihood function was maximised over d to obtain the maximum likelihood estimate for d given the data S and model M_L. The prior assumption does not influence the likelihood value and thus does not influence the maximum likelihood framework.

The goal of the Bayesian approach is to derive the posterior probability for the parameter of interest, d, under the assumed model. In this case, we know S (the data) and assume a model M_L (JC69 model on L sites together with a prior assumption on d). We want to estimate the posterior probability distribution $P(d|S, M_L)$.

Using the Bayes' rule, we have:

$$P(d|S, M_L) = \frac{P(S|d, M_L)P(d|M_L)}{P(S|M_L)} = \frac{P(S|d, M_L)P(d|M_L)}{\int_d P(S|d, M_L)P(d|M_L)\,\mathrm{d}d}. \tag{10.5}$$

This formulation highlights that the posterior is a function of both the likelihood $P(S|d, M_L)$, which is given by the JC69 substitution model (Equation (10.4)), and the prior distribution $P(d|M_L)$.

We now specify the prior distribution, $P(d|M_L)$. Suppose we know that d is less than some upper bound $d_{\max}$. Then, since d is constrained to positive values by the model, we may want to assume the following prior:

$$P(d|M_L) = \begin{cases} 1/d_{\max} & \text{for } 0 \le d \le d_{\max}, \\ 0 & \text{otherwise.} \end{cases} \tag{10.6}$$

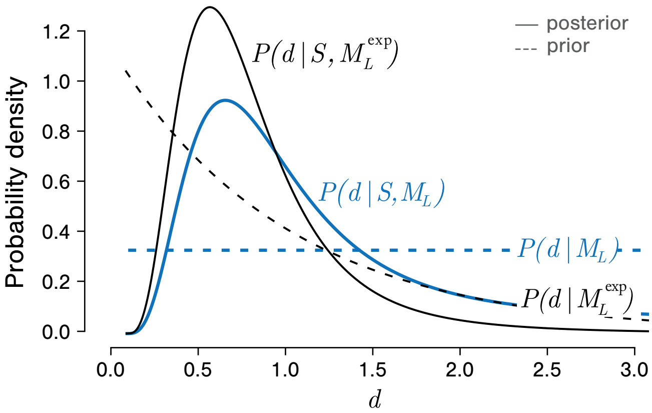


Figure 10.1: Prior and posterior distribution for genetic distance d between two sequences. Blue lines: Posterior (solid) and prior (dashed) distribution for the genetic distance d between two sequences of length $L = 10$ differing at $S = 4$ sites, assuming a JC69 model of evolution and that all values for d in the interval $(0.0, 3.0)$ are equally likely. Black lines: Posterior (solid) and prior (dashed) distribution for the same data, assuming an exponential prior for d, with the exponential truncated at 3.0.

Combining this with the likelihood (Equation (10.4)), we get:

$$P(d|S, M_L) = \begin{cases} \frac{1}{Z}\left(\frac{3}{4} - \frac{3}{4}e^{-\frac{4}{3}d}\right)^S \left(\frac{1}{4} + \frac{3}{4}e^{-\frac{4}{3}d}\right)^{L-S} & \text{for } 0 \leq d \leq d_{\max}, \\ 0 & \text{otherwise,} \end{cases} \tag{10.7}$$

where we define Z as the following integral:

$$Z = \int_0^{d_{\max}} \left(\frac{3}{4} - \frac{3}{4}e^{-\frac{4}{3}d}\right)^S \left(\frac{1}{4} + \frac{3}{4}e^{-\frac{4}{3}d}\right)^{L-S} \mathrm{d}d \tag{10.8}$$

and omit the constant binomial coefficients since they appear both in the numerator and denominator. This integral can easily be computed numerically. The resulting posterior is illustrated in Figure 10.1 for a pair of sequences of length $L = 10$ differing at $S = 4$ sites and assuming $d_{\max} = 3$, alongside the corresponding prior distribution $P(d|M_L)$.

10.1.3 Priors

As we saw at the beginning of this chapter, the appearance of $P(\theta|M)$ in Equation (10.1) is mathematically necessary to transform the likelihood of θ into a probability of θ. Thus, from the Bayesian perspective, incorporating prior/external knowledge is a basic requirement for inference.

This means that any analysis must begin with selecting prior probability distributions for the parameters of model M we want to infer. If previous inference results are available, those can be used directly (as discussed in the next section).

In other cases, we must select the most appropriate distributions given any known parameter constraints. The specific question of which prior is best to use in a given situation is difficult to answer precisely because the prior seeks to quantify information external to the data currently under investigation. As an example, consider the genetic distance inference described in the previous section. Suppose we replace our original simple constraint on d with an assertion that d is likely to be smaller rather than larger. With this assumption, we might say that $P(d|M_L^{\mathrm{exp}}) = e^{-d}$ truncated at 3 (as before). Note that we replaced M_L with M_L^{exp} to signify that our model now includes this modified assumption. This prior distribution and its effect on the posterior is shown in blue in Figure 10.1.

A common desire is to find a prior representing "no information" about a particular parameter. There are several systematic approaches to identifying such priors, such as the principle of maximum entropy for discrete variables and the method of transformation groups in the case of continuous variables (Jaynes 2003). Something to be aware of when selecting priors for continuous variables is that probability density functions are affected by changes in variables. For example, a uniform prior for some variable is not a uniform prior on the logarithm of the variable.

Since the prior is necessary for computing the posterior, the chosen prior distributions are actually a part of the model assumptions. Thus, when reporting modelling choices for an analysis, it is critical to report both the model specifying the likelihood function and the prior probability distributions. Reporting only the model specifying the likelihood is akin to reporting only a part of the assumed mathematical model, meaning that the reported results become irreproducible.

10.1.4 Incorporating additional data into a Bayesian analysis

We have mentioned that it is possible to use the posterior of a previous analysis as the prior distribution for a subsequent analysis. Here, we show that, indeed, the results are equivalent when analysing all data together or updating an analysis based on a subset of the data with the remaining data.

Consider again the problem of inferring genetic distance. Following the analysis presented in Section 10.1.2, suppose we acquire an additional pair of sequences corresponding to the same individuals analysed previously. The length of this new MSA is L', and the number of differing sites is S'. Our knowledge about the genetic distance between the sequences, in light of all data, is expressed by the new posterior $P(d|S, S', M_{L,L'})$, where $M_{L,L'}$ again represents the JC69 model, now for L and L' sites together with the prior assumption for d. Using Bayes' rule, this is:

$$P(d|S,S',M_{L,L'}) = \frac{P(S'|d,S,M_{L,L'})P(d|S,M_{L,L'})}{P(S'|S,M_{L,L'})} = \frac{P(S'|d,M_{L'})P(d|S,M_L)}{P(S'|M_{L'})}. \quad (10.9)$$

Thus, the prior for the second analysis really is the posterior for the first analysis. We may either analyse all data together or update a posterior as new data come in, using the previously inferred posterior as the prior.

10.1.5 Reporting uncertainty: credible intervals

The result of a Bayesian inference is a posterior probability distribution, which fully specifies our knowledge of parameters of interest in light of the available data and our prior assumptions. However, arbitrary probability distributions can be unwieldy to convey and may include more information than is practically necessary regarding the final state of knowledge.

Thus, results are often summarised, particularly for inferences of single parameters. Summaries can include means, variances, and other higher-order moments. In particular, it is common to summarise the results of Bayesian analyses using *credible intervals*. These regions of the parameter space carry a certain chosen percentage of the probability mass. For instance, a 95% credible interval for the parameter d from our example is an interval $[d_l, d_u]$ such that the posterior probability for d to be in that interval, given the data and model including prior information, is 0.95. One can generalise the notion of credible intervals to that of *credible sets* (applicable to discrete-valued parameters or multiple disjoint intervals on continuous parameters) and *credible regions* (for combinations of continuous parameters).

Of course, a credible interval is not unique. It is therefore common to report the so-called *highest posterior density (HPD)* interval of a parameter, defined as the smallest interval that contains the desired probability mass (Box and Tiao 1992). The *central posterior density (CPD)* interval, defined as the interval between the 0.025 and 0.975 quantiles of the posterior distribution, is also seen occasionally. Both these intervals are illustrated in Figure 10.2.

One should be aware that although credible intervals resemble confidence intervals, these are entirely different concepts. Confidence intervals characterise the distribution of possible intervals generated by all possible datasets. In quantitative terms, 95% of the estimated 95% confidence intervals contain the true (unknown) parameter. Thus, confidence intervals rely on a fixed parameter, but the boundaries are random variables. As a result, confidence intervals should never be regarded as conveying information about the possible values of the true parameter given the observed dataset. Indeed, their definition explicitly forbids this interpretation (Neyman 1937).

In contrast, credible intervals do convey this information; they are fixed intervals where the unknown parameter is the random variable: the probability that the true value is within a 95%

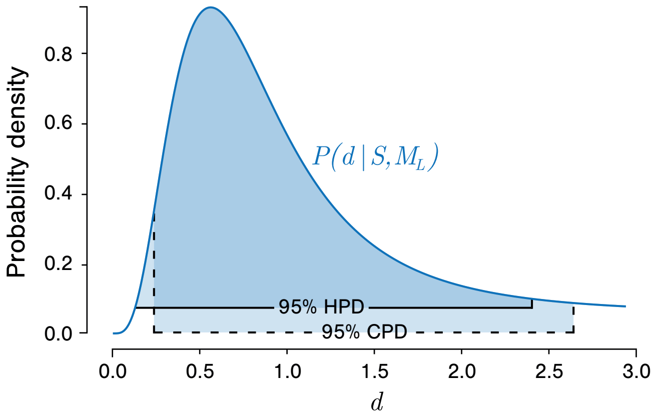


Figure 10.2: Posterior distribution for the genetic distance between two sequences of length L = 10 differing at S = 4 sites (blue line), assuming a JC69 model of evolution and a uniform prior distribution in the interval (0.0, 3.0), showing both the 95% highest posterior density (HPD) interval and the 95% central posterior density (CPD) interval.

credible interval, conditional on the observed data and prior information, is, by definition, 0.95.

10.1.6 Bayesian model selection

The Bayesian framework also provides a natural approach to judging the extent to which data support different models, which is an alternative to the approaches described in Chapter 7.

Consider two possible models, M_1 and M_2, either of which may be the true model that generated our data D. Just as for parameter inference within a particular model, we can apply Bayes theorem to the problem of selecting between different models:

$$P(M_i|D) = \frac{P(D|M_i)P(M_i)}{P(D)}. \tag{10.10}$$

This is the posterior probability that the model that generated the data D is M_i. Here i is either 1 or 2, and $P(D)$ is the result of summing the numerator over both values of i. The term $P(D|M_i)$ is the marginal likelihood introduced in Section 10.1, and here the reasoning for its name becomes clear: it is the likelihood for model M_i given the data.

With these probabilities in hand, one can judge the support for one or the other model by considering the ratio of their respective posterior probabilities:

$$\frac{P(M_1|D)}{P(M_2|D)} = \frac{P(D|M_1)}{P(D|M_2)} \cdot \frac{P(M_1)}{P(M_2)}, \tag{10.11}$$

where the denominators from Equation (10.10) cancel. Since it is common to assume equal prior support for the models under comparison, this motivates the following definition of the *Bayes factor (BF)* as the ratio of the marginal likelihoods:

$$BF = \frac{P(D|M_1)}{P(D|M_2)}. \tag{10.12}$$

Bayes factor values of less than 1 indicate that the data support M_1, while values greater than 1 indicate that the data favour M_2.

Interpretation of Bayes factors for choosing between the models is usually done on a logarithmic scale. For instance, Jeffreys (1983) suggests that a value of BF between 10^1 and $10^{1.5}$ could be considered “strong” evidence for M_1, while a value of BF greater than 10^2 could be considered “decisive” support for M_1. Similarly, a BF of less than 10^{-2} would be considered decisive support for M_2.

One challenge of using this approach to model selection is that it requires evaluating the marginal likelihood terms, which, as we discuss further in Section 10.2, can be difficult to compute in practice. However, there are several sophisticated numerical approaches to this problem, including methods such as thermodynamic integration (Lartillot and Philippe 2006) and nested sampling (Skilling 2006).

10.1.7 A Bayesian phylogenetics and phylodynamics framework

In the previous sections, we introduced the Bayesian framework. In what follows, we employ this framework in a joint phylogenetic and phylodynamic analysis.

More concretely, suppose we have an MSA A of n sequences, which may have been collected at different times. We assume these sequences evolved from an unknown common ancestor along a time tree $\mathcal{T}$ according to a substitution model with rate matrix Q. Furthermore, we assume that the time tree $\mathcal{T}$ itself is the product of a population dynamic process parameterised by η. In the case of a birth-death model, η would be the birth, death, and sampling rate parameters together with the start time of the process, while in the case of a coalescent model, η would be the effective population size and the generation time. Again, we use M to formally denote the combination of our model assumptions, specifying the likelihood and any prior information on the model parameters.

To formulate a Bayesian analysis, we first express what we want to learn from the analysis in terms of a conditional probability distribution. Here, we want to learn about all the unknown aspects of the past evolutionary process (tree, substitution model parameters, tree-generating model parameters) conditional on our data A and our model M, $P(\mathcal{T}, Q, \eta | A, M)$.

Applying Bayes' rule, this can be written as:

$$P(\mathcal{T}, Q, \eta | A, M) = \frac{P(A | \mathcal{T}, Q, \eta, M) P(\mathcal{T}, Q, \eta | M)}{P(A | M)}. \tag{10.13}$$

The definition of conditional probability (Section 1.3.1) allows us to write $P(\mathcal{T}, Q, \eta | M) = P(\mathcal{T} | Q, \eta, M) P(Q, \eta | M)$, so the equation becomes:

$$P(\mathcal{T}, Q, \eta | A, M) = \frac{P(A | \mathcal{T}, Q, \eta, M) P(\mathcal{T} | Q, \eta, M) P(Q, \eta | M)}{P(A | M)}. \tag{10.14}$$

This expression of the posterior distribution is completely general.

We make the following model assumptions in most phylogenetic and phylodynamic analyses. First, we assume that the MSA depends only on the substitution process and the tree, and it has no dependence on the phylodynamic parameters η. Thus $P(A | \mathcal{T}, Q, \eta, M) = P(A | \mathcal{T}, Q, M)$.

Second, the tree is a product only of the phylodynamic process, and thus $P(\mathcal{T} | Q, \eta, M) = P(\mathcal{T} | \eta, M)$.

Finally, we assume that our prior information relating to Q is independent of our prior information relating to η, meaning we can write $P(Q, \eta | M) = P(Q | M) P(\eta | M)$.

With these assumptions, the expression for the posterior distribution is:

$$P(\mathcal{T}, Q, \eta | A, M) = \frac{P(A | \mathcal{T}, Q, M) P(\mathcal{T} | \eta, M) P(Q | M) P(\eta | M)}{P(A | M)}, \tag{10.15}$$

and it involves terms that we already discussed: the phylogenetic likelihood $P(A | \mathcal{T}, Q, M)$, the tree probability density or tree prior $P(\mathcal{T} | \eta, M)$, the prior distributions $P(Q | M)$ and $P(\eta | M)$, and the marginal likelihood $P(A | M)$.

Note that in the last section, we called the expression $P(\mathcal{T} | \eta, M)$ — which is the probability density of the tree — the phylodynamic likelihood, emphasising a function in η where $\mathcal{T}$ was the data (and optimising over η resulted in maximum likelihood estimates). Now, the data are A, and $P(\mathcal{T} | \eta, M)$ is part of the prior, also called tree prior.

Bayesian phylogenetic and phylodynamic inference methods aim to determine the posterior distribution $P(\mathcal{T}, Q, \eta | A, M)$. However, despite the apparent simplicity of this problem, it presents a major challenge because

(i) the probability distribution involves a highly multi-dimensional state space (the space of possible combinations of $\mathcal{T}$, Q, and η for an MSA of a given size), and

(ii) the distribution itself often has a complex structure with multiple peaks for analyses of practical interest; this complex structure, due to the combination of discrete tree topologies and continuous branch length parameters, has been investigated numerically by Whidden and Matsen (2015).

In particular, the large dimensionality of the state space means that the denominator on the right-hand side, the marginal likelihood, is virtually impossible to evaluate directly, as we did in our toy example (Section 10.1.2). The difficulty can be seen more clearly if we expand it in terms of the distributions found in the numerator:

$$P(A|M) = \int_{\mathcal{T},Q,\eta} P(A|\mathcal{T},Q,M)P(\mathcal{T}|\eta,M)P(Q|M)P(\eta|M)\,\mathrm{d}\mathcal{T}\,\mathrm{d}Q\,\mathrm{d}\eta. \tag{10.16}$$

Directly evaluating this integral would involve summing over every possible combination of $\mathcal{T}$, Q, and η. Not only would this require integrating over many continuous parameters, but it would also require evaluating the probability of every single possible tree $\mathcal{T}$ relating the sequences in the MSA. As we have noted previously (see Section 6.2.3), the number of such trees is extremely large, even for small numbers of sequences, making this integration highly impractical.

The following section introduces a framework for determining the posterior distribution without evaluating the marginal likelihood.

10.2 Markov chain Monte-Carlo (MCMC) sampling

In Bayesian analyses, the marginal likelihood is often impossible to calculate directly. This is particularly true for phylogenetic applications that, combined with many parameters, have a large state space that includes all possible phylogenetic trees.

The marginal likelihood $P(D|M)$ does not depend on the model parameter θ. Thus, provided we can compute the likelihood and the prior for a particular combination of parameter values, we can still compute the posterior probability of θ, given the data D, up to a multiplicative constant (namely up to the marginal likelihood):

$$P(\theta|D,M) = \frac{1}{P(D|M)} \cdot P(D|\theta,M)P(\theta|M) = \frac{1}{Z}f(\theta). \tag{10.17}$$

However, not knowing the normalising constant Z prohibits us from judging whether some parameter value θ' is probable or not. Imagine that for $\theta = \theta'$, we compute $f(\theta') = P(D|\theta = \theta', M)P(\theta = \theta'|M)$. The result of this product is proportional to the posterior probability density $P(\theta = \theta'|M, P)$ shown in Figure 10.3. However, since we do not know the normalisation, we cannot know whether the probability density of θ' is large or small.

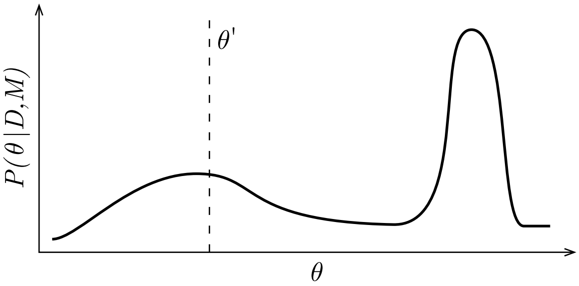


Figure 10.3: An example posterior distribution. This illustrates the necessity of knowing the normalisation of the distribution. Without such a normalisation, there is no way of knowing that the posterior probability density for values in the vicinity of $\theta = \theta'$ is, in fact, relatively low due to the large peak elsewhere.

To solve this problem, we will use a technique that compares posterior probabilities and thus does not need to calculate the normalising constant directly. Bayesian inference commonly uses *Monte Carlo algorithms*[2] to characterise posterior probability distributions. Instead of directly evaluating the probability density of the particular parameter values, these algorithms produce random samples from the posterior distribution using an algorithm that does not require evaluating the normalising constant. For example, suppose we draw samples $\theta^{(1)}, \theta^{(2)}, \ldots$ from $P(\theta|D,M)$. We can then compute the mean of the posterior by using the fact that the average of the samples approaches the mean of the distribution in the limit of a large number of samples:

$$\mathrm{E}(\theta) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \theta^{(i)} = \int \theta P(\theta|D,M)\,\mathrm{d}\theta. \tag{10.18}$$

We can compute the variance in a similar way:

$$\mathrm{Var}(\theta) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (\theta^{(i)} - \mathrm{E}(\theta))^2 = \int (\theta - \mathrm{E}(\theta))^2 P(\theta|D,M)\,\mathrm{d}\theta. \tag{10.19}$$

In fact, with enough of these samples, it is possible to answer any question regarding the shape of the underlying distribution.

The particular Monte Carlo algorithm used frequently in phylogenetics and many other applications of Bayesian statistics is the so-called *Markov chain Monte Carlo (MCMC)* algorithm. It is also known as the Metropolis-Hastings algorithm, as its current general form was developed in a pair of papers by Nicholas Metropolis and two husband and wife teams — Arianna and Marshall Rosenbluth, and Augusta and Edward Teller — in 1953 (Metropolis et al. 1953). Hastings (1970) further elaborated on this method. The original application of

[2] Named in honour of a famous casino in Monte Carlo.

the algorithm was to solve problems in statistical physics, but today, it is used in a wide array of fields.

10.2.1 MCMC algorithm

Consider a parameter θ distributed according to some distribution $\pi(\theta)$ which we want to characterise. Imagine the space of all possible parameter values θ in a particular model. A point in this space, which we refer to as a state, precisely identifies one configuration of parameter values. The goal of MCMC is to characterise a probability distribution $\pi(\theta)$ over this state space. In our context, this distribution is the posterior probability for θ, $P(\theta|D, M)$.

The MCMC algorithm approaches this goal by constructing a random walk through the parameter space such that the frequency with which each state is visited is proportional to the posterior probability of that state. This random walk is a series of "steps" through this space that fulfil the properties of a Markov chain: each step is independent of the previous steps and only depends on the current position. We refer to this walk as the "chain".

Specifically, we define a Markov chain $\theta^{(1)}, \theta^{(2)}, \ldots$ by setting the initial value of the parameters $\theta^{(1)}$ arbitrarily, and then for each subsequent step applying the following rules:

(i) choose a new state θ' from some (easy to sample) distribution $q(\theta'|\theta^{(i)})$, where i is the index of the previous step;

(ii) accept the new state with probability $\alpha(\theta'|\theta^{(i)})$ and set $\theta^{(i+1)} \leftarrow \theta'$, otherwise reject the new state and set $\theta^{(i+1)} \leftarrow \theta^{(i)}$.

Our goal is to define the functions q and α such that the chain possesses a stationary distribution, and this distribution is the target distribution $\pi(\theta)$. That is, after many steps, we want the probability of finding the chain in state θ to be $\pi(\theta)$.

To meet this goal, we firstly require the *proposal distribution* $q(\theta'|\theta)$ to be chosen such that the chain can reach every possible value of the parameter θ when considering only those for which $\pi(\theta) > 0$. Note that states only need to be reachable *eventually*, not necessarily in a single step. The chain thus defined has the irreducibility property described in Section 5.2.4.

Secondly, we require that the chain fulfils the detailed balance condition (Equation (5.42)):

$$\pi(\theta)\alpha(\theta'|\theta)q(\theta'|\theta) = \pi(\theta')\alpha(\theta|\theta')q(\theta|\theta'). \tag{10.20}$$

This implies that the chain has a stationary distribution and that $\pi(\theta)$ is this stationary distribution (Lemma 5.3.2). We can ensure this detailed balance condition is satisfied by defining the function $\alpha(y|x)$ in the following way:

$$\alpha(y|x) = \min\left(1, \frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}\right). \tag{10.21}$$

That this satisfies Equation (10.20) can be seen by substitution.

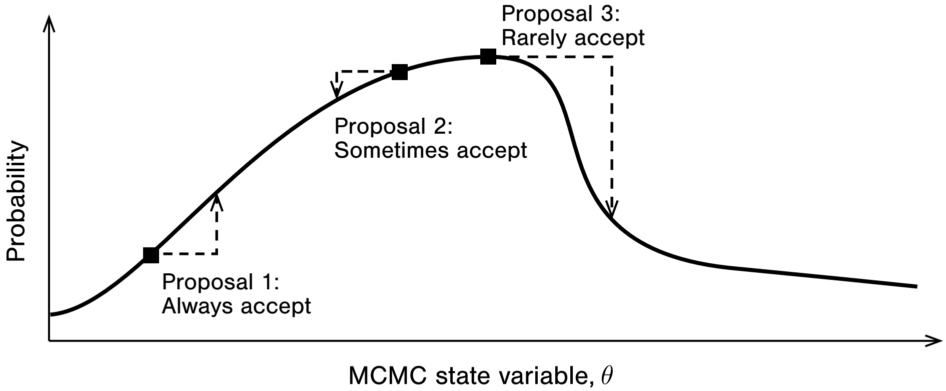


Figure 10.4: Illustration of three different proposals during an MCMC analysis. In Proposal 1, the proposed state (pointed to by the dashed arrow) has a higher probability than the original state (solid square) and will thus always be accepted. In Proposal 2, the proposed state has a slightly lower probability than the previous state, meaning the move will be frequently — but not always — accepted. For Proposal 3, the proposed state has a much lower probability than the original state and will thus be rarely accepted. Note that this picture is correct only when the proposal satisfies the symmetry $q(\theta'|\theta) = q(\theta|\theta')$.

Because the acceptance probability $\alpha(\theta'|\theta)$ is the only place that the target distribution (that is, the posterior probability distribution) enters the algorithm, the fact that it depends only on the ratio between the probability distribution evaluated at θ and θ' means that we do not need to know the marginal likelihood to use the algorithm.

This procedure ensures the chain will make more steps toward higher probability states. Moreover, in the limit of an infinite number of steps, the relative number of visits to a particular state will precisely equal the posterior probability of that parameter configuration.

Looking at Equation (10.21) can give some insight into how this works. When the state proposal distribution is symmetric such that $q(\theta'|\theta) = q(\theta|\theta')$, we can see that the acceptance probability $\alpha(\theta'|\theta)$ is 1 whenever $\pi(\theta') \geq \pi(\theta)$. As the posterior probability of the proposed state drops below that of the current state, the acceptance probability also drops according to the ratio of the probability of the proposed state relative to that of the current state. Thus, the chain is biased toward high-probability states while continuing to explore lower-probability states.

Figure 10.4 roughly illustrates the properties of the MCMC random walk as they apply to sampling a one-dimensional probability distribution.

Box 31: Glossary of MCMC terms

MCMC algorithm: a random walk through a parameter state space such that the frequency with which each parameter configuration is visited is proportional to a targeted posterior probability distribution.

State: a single realization of the parameters whose posterior distribution we seek to quantify.

State space: all possible parameter value combinations; the domain of the posterior distribution which the MCMC algorithm aims to characterise.

Step/iteration: the index of a single state visited by an MCMC algorithm.

Chain: a sequential list of states produced by the MCMC algorithm.

Proposal distribution: the probability distribution that governs how the state at the next step in an MCMC chain is generated from the current state.

Acceptance probability: the probability that a state drawn from the proposal distribution is accepted as the next state of the chain.

Convergence: "equilibration" of the MCMC chain to the target posterior distribution. Roughly, the point at which one can say that the chain's current state represents a draw from the posterior. Before this point, the starting state of the chain still had an influence on the distribution.

Burn-in: the number of steps required for a chain to converge. The burn-in is removed from the output.

Autocorrelation: statistical correlation between states at nearby steps in a chain.

Effective sample size (ESS): an estimate of the effective number of independent draws from the posterior represented by a given chain, taking into account the autocorrelation between states in the chain. The ESS is usually much smaller than the chain length, and results should only be taken seriously when some minimum ESS has been achieved.

Mixing rate: a colloquial term used to refer to the speed at which an MCMC chain converges and acquires effectively independent samples. It is affected by many things, including the proposal distribution used and the posterior distribution itself. When this rate is slow, we often say that the chain is *mixing slowly* or *mixing poorly*.

Log: a record of the states visited by an MCMC algorithm. Usually, only a fraction of visited states are included, for example, 1 logged state for every 1 000 steps.

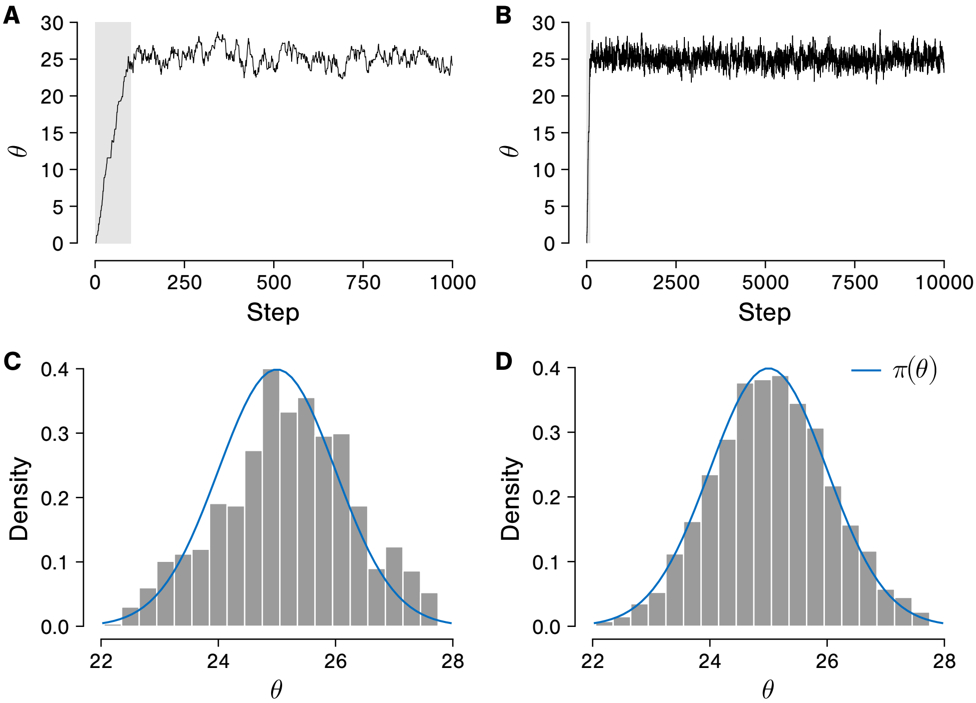


Figure 10.5: **A** Plot showing the state trajectory of the MCMC as a function of the step number, with the burn-in period highlighted in grey. **C** Histogram showing the relative frequency of the states visited by the chain (after burn-in) and the target probability distribution (blue). **B, D** State trajectory and histogram of a longer chain. Note the close correspondence between the relative frequencies and the target distribution.

10.2.2 Toy example

We now illustrate the MCMC algorithm in a toy example. In this example, we want to produce samples from the target distribution $\pi(\theta) \propto \exp\left({}^{-(\theta-25)^2}/{}_2\right)$. To do this, we define a uniform proposal distribution with the proposal density defined such that $q(\theta'|\theta) = 1/2W$ when $|\theta - \theta'| < W$ and 0 otherwise, with W being some fixed positive number. Again, recall that this choice is highly arbitrary — all that is required is for a chain of such steps to be capable of reaching every value of θ.

The acceptance probability function then becomes:

$$\alpha(\theta'|\theta) = \min\left[1, \frac{\pi(\theta')q(\theta|\theta')}{\pi(\theta)q(\theta'|\theta)}\right] \tag{10.22}$$

$$= \min\left[1, \frac{\exp\left(^{-(\theta'-25)^2}/_2\right)}{\exp\left(^{-(\theta-25)^2}/_2\right)}\right], \tag{10.23}$$

where the proposal probabilities cancel out since, in this case, $q(\theta'|\theta) = q(\theta|\theta')$.

The result of running this chain for 1 000 steps, starting from an initial state $\theta^{(1)} = 0$ and using a proposal distribution with $W = 1$, is shown in Figure 10.5 **A**. Here, we see that the chain initially heads toward the peak of the probability distribution at $\theta = 25$, then fluctuates around this maximum. Clearly, the choice of the initial state influences the first part of the chain very heavily. This initial period (highlighted in grey in the figure) of the chain where the initial state continues to influence the result is known as the *burn-in*, and the first step in analysing the results of an MCMC chain is to identify and discard this burn-in period (for the definition of these and other MCMC-related terms, see Box 31 on page 315).

Figure 10.5 **C** shows a histogram of the states visited by the chain (after removing the burn-in period), compared with the true density $\pi(\theta)$ we are aiming to sample. We see that there is already a close correspondence between the sampled histogram and target distribution, even after only 1000 steps.

Figure 10.5 **B** and Figure 10.5 **D** show the chain and corresponding histogram after 10 000 steps. Here, we see almost perfect agreement between the histogram and the target distribution.

10.2.3 Mixing rate and effective sample size

As shown above, an important property of the MCMC algorithm is that any proposal density $q(\theta'|\theta)$ that is able to explore the whole state space (that is, makes it possible to eventually reach any state with enough steps) can be used to produce a correct MCMC algorithm. By "correct," we mean that the relative frequencies of visits to each state will eventually match the target distribution in the limit of an infinite number of steps. However, practical applications of MCMC can only ever involve a finite number of steps. Thus, we are interested in the number of steps required to characterise the target distribution well.

Continuing with the example from the previous section, we consider two different scenarios. In the first scenario, we modify the proposal density so that the proposal window W takes a much smaller value of 0.1. This means that each step can only change θ by a tiny amount.

Figure 10.6 **A** shows the state trajectory resulting from 10 000 steps with this modified proposal distribution. We can see that the burn-in process takes much longer to complete (around 2 000 steps, compared to just around 100 for the $W = 1$ proposal). Furthermore, even after

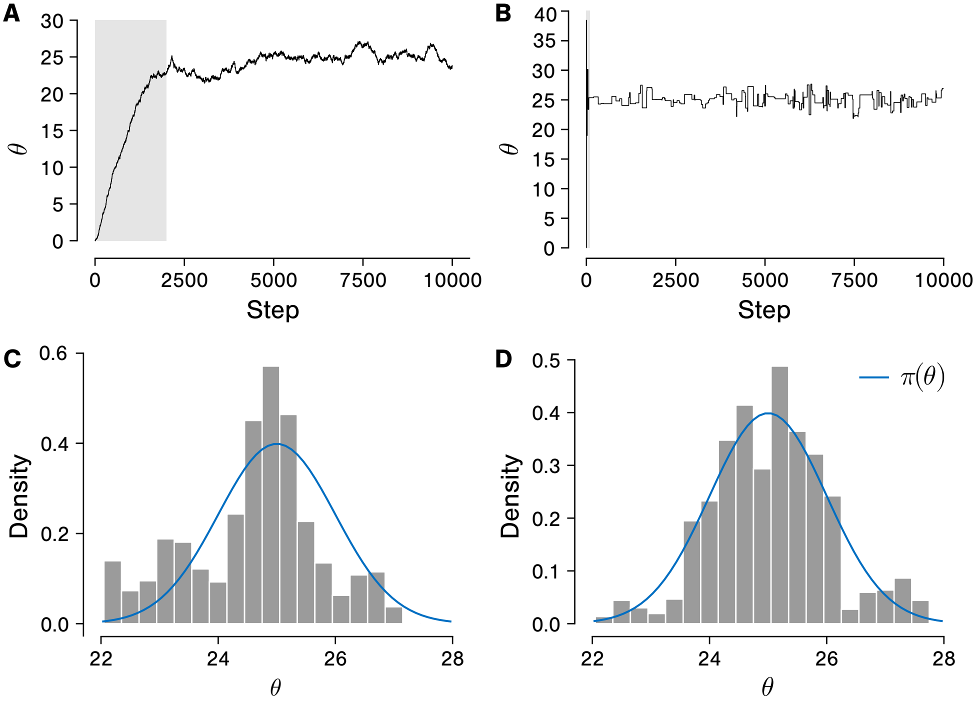


Figure 10.6: **A, C** State trajectory and histogram of an MCMC chain as in Figure 10.5 but reduced proposal window $W = 0.1$ (instead of $W = 1$). This proposal is not bold enough, leading to a poorly mixing chain with few effectively independent samples. **B, D** State trajectory and histogram produced using a very large proposal window $W = 100$. This proposal is far too bold, which also leads to poor mixing.

the chain has reached the peak of the target distribution, the fluctuations about this point are much slower. We say such a chain is *mixing slowly* or *mixing poorly*. Such slow mixing reduces the quality of the characterisation of the target density, as is evident from the associated histogram in Figure 10.6 **B**, which displays a much weaker correspondence to the true distribution.

We now consider a second scenario in which the proposal density is modified so that the proposal window W has a much larger value of 100. This means that each step can produce a very wide range of values.

As we see in Figure 10.6 **C**, this situation is also problematic. While the burn-in period is extremely fast, the chain appears “jagged”, and the associated histogram in Figure 10.6 **D** is again quite far from the target distribution. What we see is the result of the proposal being too “bold” for this target density. The proposal distribution often produces states that are very different from the current state, meaning that if the current state is close to the peak of the target distribution, it is highly likely that the proposed state will be very far away from this peak and will thus have a very low probability. Such proposals are rarely, if ever, accepted, meaning that for many steps in this chain, the state remains the same, leading to long runs of identical states and hence the step-like appearance of the state trajectory. Thus, this scenario also produces a slow mixing chain.

Some MCMC practitioners use the concept of *effective sample size (ESS)* to quantify the information that a given chain carries about the target distribution. The effective sample size is loosely defined as the number of effectively independent samples from the target distribution contained in a given MCMC chain. Because each step in the MCMC chain produces a new state in a manner that is conditional on the current state, nearby states tend to be highly correlated, meaning the ESS will always be less than the total number of steps. This is true even when the proposal distribution $q(\theta'|\theta)$ is independent of the previous state since the possibility of rejection can always induce autocorrelation. In the example above, we saw the ESS can become very low if the width of the proposal distribution is too narrow. We refer the interested reader to Magee et al. (2023) for a good overview of ESS and autocorrelation in a Bayesian phylogenetic MCMC context.

In practice, users will generally set a predefined ESS threshold to decide whether the algorithm has run for enough steps and sampled sufficiently from the target distribution or should run for more steps.

The example scenarios illustrated that, even in the simple case of a single-parameter model, the choice of proposal distribution can strongly influence the number of steps required for the samples to characterise the target distribution usefully. This influence is even more pronounced when performing MCMC on the large multivariate problems commonly encountered in Bayesian phylogenetics and phylodynamics.

10.2.4 MCMC for phylogenetics and phylodynamics

Markov chain Monte Carlo is used in Bayesian phylogenetics in order to produce samples from and thus characterise the posterior distribution given by Equation (10.15), $P(\mathcal{T}, Q, \eta|A, M)$.

As we have seen, the MCMC algorithm relies on appropriate proposal distributions that allow the resulting chain to explore the parameter space efficiently. In phylogenetic and phylodynamic applications, new values for the continuous parameters (e.g. Q, η) are usually proposed by scaling the current values, whereas new tree topologies ($\mathcal{T}$) are proposed via the moves presented in Chapter 6: nearest-neighbour interchange (NNI), subtree pruning and regrafting (SPR) and tree bisection and reconnection (TBR).

The output of such an MCMC algorithm is a series of samples from the parameter space. Each sample contains the tree topology, associated branch lengths, and the values of all continuous parameters from the evolutionary (Q) and the tree-generating (η) models.

Figure 10.7 illustrates the posterior distribution for the phylogenetic time tree relating a number of H1N1 influenza A virus sequences from the 2009 Swine Flu pandemic collected from publicly-available data by Hedge, Lycett and Rambaut (2013), as sampled using MCMC. In this figure, sampled trees are drawn on top of one another to illustrate the uncertainty in the inferred tree remaining after the data are taken into account.

Together with samples of the trees, we obtain samples of Q and η. Thus, we naturally perform a joint phylogenetic (leading to the tree and the substitution rate matrix) and phylodynamic (leading to η) analysis. The parameter estimates for η, in particular, take the tree uncertainty into account.

There are a large number of software tools available that implement MCMC for Bayesian phylogenetic applications. Some of the most popular include:

- BEAST (`https://beast.community/`) (Drummond and Rambaut 2007) and BEAST2 (`https://www.beast2.org/`) (Bouckaert et al. 2014; Bouckaert et al. 2019),
- MrBayes (`https://nbisweden.github.io/MrBayes/`) (Huelsenbeck and Ronquist 2001), and
- RevBayes (`https://revbayes.github.io/`) (Höhna et al. 2016).

10.2.5 Time complexity of phylogenetic MCMC algorithms

We have seen several approaches to inferring phylogenetic trees in this book, ranging from extremely simple techniques, such as the UPGMA algorithm, to the sophisticated Bayesian techniques we have discussed in this chapter. For each of the previous algorithms, we have discussed how the runtime of the algorithm scales as a function of the size of the tree (that is, the number of leaves). We now ask: what is the time complexity of MCMC-based Bayesian phylogenetic inference algorithms?

For MCMC, a critical quantity is the burn-in time. As already discussed in Section 10.2.3, this is the approximate number of steps required for a chain to reach its stationary distribution (the posterior) after being initialised using a given starting state. This is also related to the effective sample size since the ESS is also determined by the number of steps necessary for the chain to “forget” about an earlier state.

It can be shown that the burn-in time for simple MCMC algorithms in d-dimensional continuous real parameter spaces sampled using simplistic multivariate normal proposal distributions grows linearly in d (Roberts and Rosenthal 2016). However, the phylogenetic situation is far more complex, involving both continuous and highly structured discrete parameters. Furthermore, different sets of sequence data can produce posterior distributions with very different qualitative characteristics. For instance, in a comprehensive empirical study, Whidden and Matsen (2015) demonstrated that some phylogenetic posteriors can be strongly multi-modal and thus difficult to sample using MCMC, while others can appear almost uni-modal. This

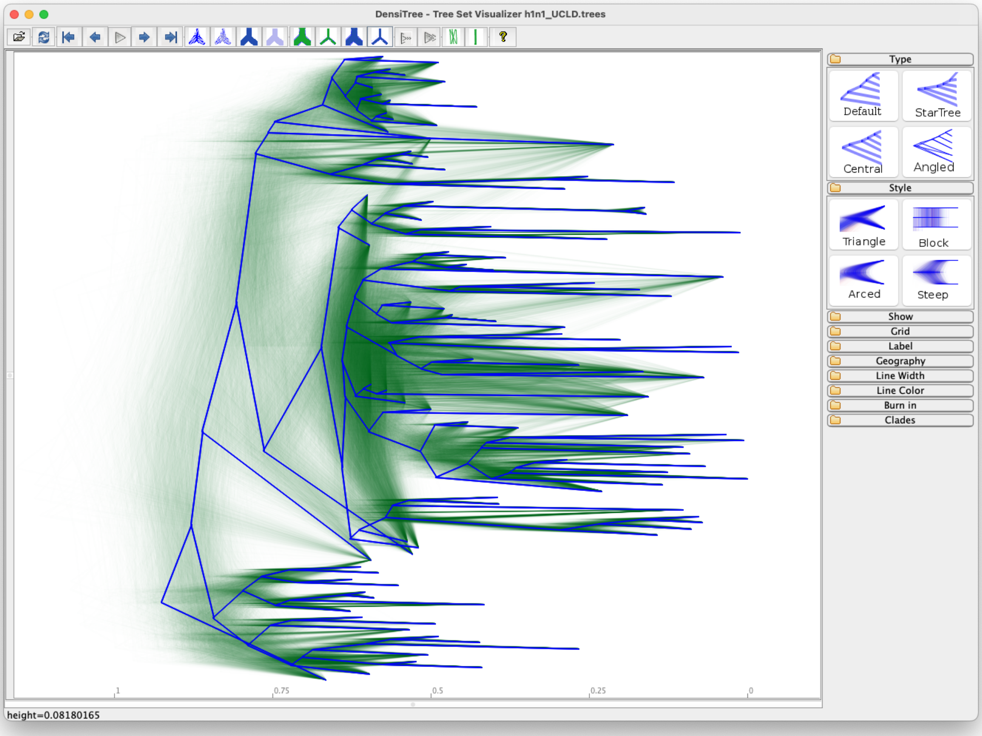


Figure 10.7: Posterior distribution of the phylogenetic tree relating a set of H1N1 influenza A virus sequences, visualised in DensiTree (Bouckaert and Heled 2014) by drawing the time trees sampled by the MCMC chain on top of one another in green. The highest probability topology is depicted in blue and can be regarded as the mode of the tree distribution.

results in a wide variation in the number of steps necessary for characterisation across datasets.

For these reasons, it is clear that the asymptotic runtime estimates from simpler models should not be applied to the phylogenetic case and that the question, in general, remains open.

10.3 Applications

We close this chapter by highlighting several studies using Bayesian phylogenetic and phylodynamic inferences for macroevolutionary and epidemiological applications. For the presented studies, we outline the main ideas and results. We refer the reader to the cited papers for information about the assumed prior distributions and further details of the analyses. All studies used the software platform BEAST2 (Bouckaert et al. 2014; Bouckaert et al. 2019).

10.3.1 Epidemiology: The Ebola epidemic in West Africa, 2013-2016

As we saw in Chapter 9, the Ebola epidemic in West Africa 2013-2016 was studied using phylogenetic and phylodynamic methods. In particular, we discussed that the maximum likelihood estimates for the birth and death rates of the epidemic on a fixed phylogenetic tree were obtained by Stadler et al. (2014), giving an estimate of the basic reproductive number $R_0 = 1.34$ (CI $[1.12, 1.55]$). Applying Bayesian MCMC inference takes into account the phylogenetic uncertainty when estimating R_0, leading also to less certain R_0 estimates (median 1.65, 95% HPD interval $[1.02, 2.70]$ (Stadler et al. 2014)).

Finally, using the birth-death skyline approach in a Bayesian MCMC setting, the effective reproductive number through time in that Ebola outbreak was estimated in du Plessis (2016), showing that the effective reproductive number started declining drastically in the second half of May 2014 (Figure 10.8), well before the epidemic became a WHO “public health emergency of international concern” in early August 2014 (Check Hayden 2014).

10.3.2 Epidemiology: The SARS-CoV-2 pandemic

The evolutionary dynamics of the SARS-CoV-2 pandemic have been studied extensively using Bayesian phylogenetic and phylodynamic methods. Indeed, the study discussed earlier in Section 9.1.6 describing per-country estimation of R_0 from genomic data was conducted as a Bayesian phylodynamic study.

Another Bayesian analysis was conducted as part of the study by Nadeau et al. (2023), in which over 5000 SARS-CoV-2 virus genomes collected in Switzerland during the year 2020 were grouped into small clusters representing independent Swiss transmission chains. A large birth-death phylodynamic analysis was conducted in which the phylogenetic tree of every cluster, together with overall rates of sampling and transmission, allowing for cluster-specific transmission rate variation, were jointly inferred using Bayesian MCMC.

These inference results were used to identify genomic evidence for the efficacy of Swiss contact tracing measures, which were assumed to produce a cluster-specific reduction in transmission rate shortly after detecting the first sample in each cluster. Figure 10.9 shows the relevant figure from the paper, which illustrates the posterior distributions of cluster-specific transmission rate damping at different times throughout the year. Based on the genomic sequence data, the study suggests that contact tracing was effective in the summer of 2020 — slowing transmission by around 50% — but was overwhelmed in the fall of 2020.

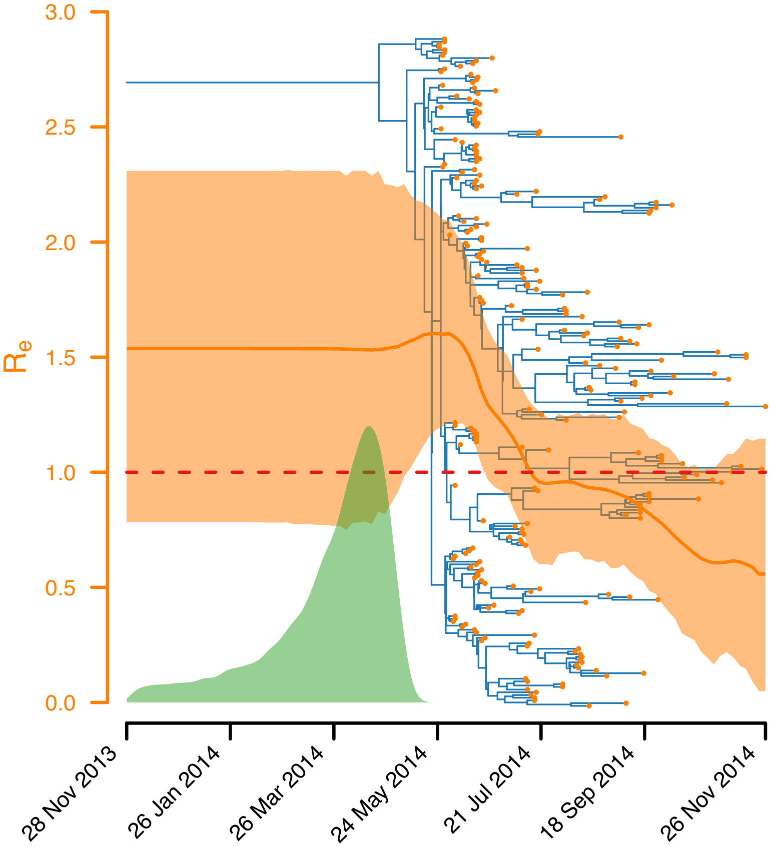


Figure 10.8: Bayesian analysis of Ebola sequences from West Africa outbreak 2013-2016. Phylogenetic tree (blue), posterior distribution of the start of the epidemic (green) and median with 95% HPDs of the effective reproductive number (orange) obtained from a BEAST2 analysis. Figure adapted from du Plessis (2016).

10.3.3 Macroevolution: Phylogeny of penguins

This example illustrates using Bayesian inference to study macroevolution, in this case, the evolution of penguins (Gavryushkina et al. 2017). The analysis used the DNA sequences of extant species and the dates and morphological characteristics of several fossils.

The morphological characteristics were included as sequences of morphological characters produced by domain experts analysing each fossil. These morphological sequences were then organised into a table resembling an MSA, allowing their evolution to be modelled as the result of a continuous-time Markov chain process analogous to the processes used to model

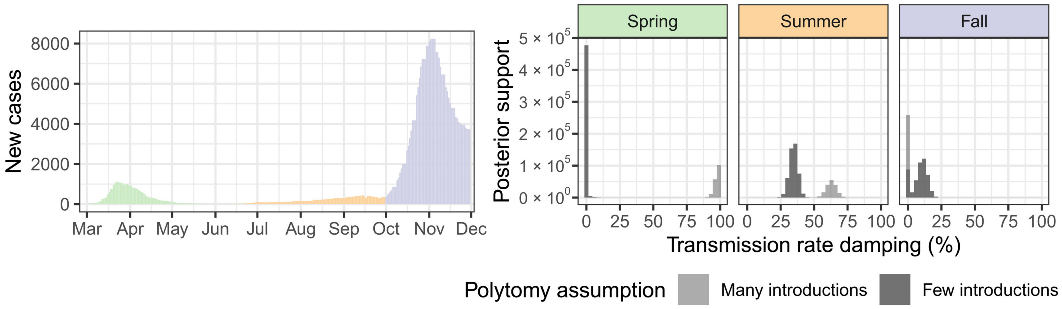


Figure 10.9: Bayesian analysis of SARS-CoV-2 sequences from Switzerland collected in 2020 performed in BEAST2. The panel on the left shows the number of recorded cases throughout the year, with the colours indicating the season. The panel on the right shows the posterior distributions of the clade-specific transmission-rate damping factors for each season, with higher factors implying more transmission damping. These results indicate that contact tracing efforts were most effective in the summer — since the dampening factor was highest in summer — but were overwhelmed in the fall. The two different "polytomy assumptions" relate to two different approaches to clustering sequences into transmission chains; both sets of results were included as a robustness check. Figure adapted from Nadeau et al. (2023).

genetic sequence evolution.

A Bayesian MCMC analysis was performed using morphological and molecular evolution models and a birth-death model generating the tree. In the birth-death model, sampling was performed through time (giving rise to fossils; the removal probability was zero) and at present (giving rise to extant penguins). This specific setting of sampling in the birth-death model is also called the *fossilised birth-death process* (Heath, Huelsenbeck and Stadler 2014).

The phylogeny obtained is shown in Figure 10.10. In particular, using all data in a joint analysis suggested that the most recent common ancestor of the extant penguins is much younger than previously thought. Note that in this analysis, the posterior distribution of the phylodynamic model parameters was not explored further; however, these parameters are an essential component of the analysis even when we are merely interested in the phylogenetic tree.

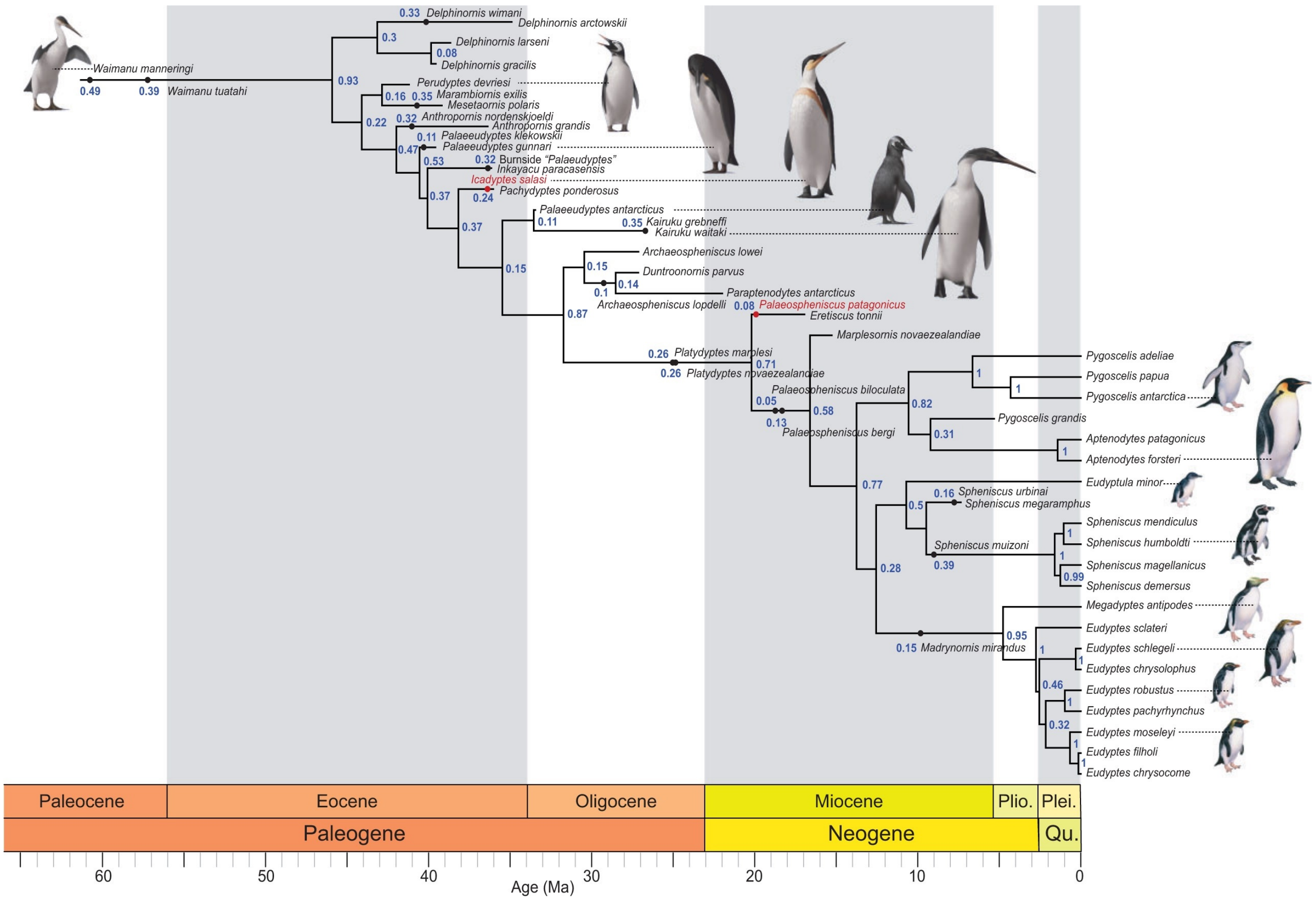


Figure 10.10: Penguin phylogeny obtained from fossil dates, morphology, and extant species sequence data. Analysis was performed in BEAST2 using a special birth-death model, the fossilised birth-death model (Heath, Huelsenbeck and Stadler 2014). Figure adapted from and methods published in Gavryushkina et al. (2017).

11 Phylogenetic networks

This book has so far focused on scenarios where evolution occurs in a tree-like fashion (but see Chapter 4 where we assumed that sites are completely independent/unlinked due to, for example, recombination). A tree means that all individuals descend from a single ancestor, and each individual has precisely one parent from whom it inherits all genetic material. The phylogenetic and phylodynamic models we have discussed so far have this tree assumption built into their cores.

While the tree-based view of evolution is tremendously useful, there are many situations where it is false. Eukaryotic organisms almost always reproduce sexually; thus, all individuals have two parents. While individual alleles often come from just one of these parents, recombination ensures that sequences may share their ancestry between multiple parents. Bacteria may exchange and acquire new genetic material through various mechanisms such as contact-dependent plasmid transfer, phage-mediated transduction, and transformation via DNA uptake from the environment (Thomas and Nielsen 2005). Virus ancestry may be non-tree-like due to recombination and reassortment (Pérez-Losada et al. 2015).

Non-tree-like processes do not only occur within populations of individuals, single cells, or virions. On the macroevolutionary level, multiple parental species may hybridise to form a single descendant species (Mallet 2007) or species may exchange genetic material through horizontal gene transfer (Soucy, Huang and Gogarten 2015).

Due to the prevalence and importance of these events, it is now sometimes suggested that instead of talking about the “tree of life” we might instead refer to a “network of life” (e.g. see Ragan, McInerney and Lake (2009) and Doolittle (1999)). This chapter will focus on several ways these more general evolutionary processes can be accounted for in phylogenetic and phylodynamic analyses. This chapter can be seen to fall between the previous chapters on trees where none of these network events occurred, and Chapter 4, where the network processes were so dominant that we could treat all data points as independent.

11.1 Mathematics of phylogenetic networks

Mathematically, *phylogenetic networks* are a generalisation of phylogenetic trees as defined in Section 6.2. Generally, we consider unrooted or rooted *binary networks*. As in trees, nodes in binary *unrooted phylogenetic networks* may be either degree-1 or degree-3, with degree-1 nodes referred to as leaves or tips and degree-3 nodes referred to as internal nodes. *Rooted networks* additionally have a single degree-2 node called the root.

The major difference is that, unlike trees, phylogenetic networks may have cycles. This means there may be more than one sequence of branches connecting a given pair of nodes in the network.

In a rooted tree, each degree-3 node has one ancestor branch and two descendant branches. In a rooted network, a degree-3 node may again have one ancestor and two descendant branches (*coalescence nodes*); or it may have two ancestor branches and one descendant branch (*reticulation nodes*).

An important consequence of allowing cycles is that the number of possible distinct networks on n leaves is infinite, while the number of distinct trees on n tips is finite (Section 6.2.3). To see this, consider the rooted phylogenetic tree shown in Figure 11.1 **A**. This tree can be transformed into a rooted phylogenetic network by adding one reticulation node, one coalescence node, and an edge connecting them, as shown in Figure 11.1 **B**. We can keep adding edges, as shown in Figure 11.1 **C** and **D**. Since each of these phylogenetic networks is distinct and valid, and since we can always add an edge to form a new network, the number of possible unique phylogenetic network relationships between a given number of leaves is trivially infinite. This implies that introducing cycles infinitely expands the already large space of possible phylogenies, potentially making the inference problem even more challenging.

More details on the mathematical properties of phylogenetic networks and algorithmic approaches can be found in Huson, Rupp and Scornavacca (2010).

In what follows, we discuss phylogenetic networks in the context of different evolutionary processes.

11.2 Recombination networks

Phylogenetic networks are used to study the effects of recombination. On a general level, recombination describes the combination of “parental” genetic sequences to produce a “child” sequence. Thus, a strand of DNA of a child contains information from the DNA strands of both parents. Depending on the exact molecular mechanism, recombination can result in either

(i) inserting non-homologous genetic material into a genome, or

(ii) replacing a portion of genetic material with homologous material.

The latter is known as *homologous recombination* and is the exclusive focus of this chapter due to its effect on phylogenetic reconstruction from homologous sequence alignments. Homologous recombination of some form occurs in many parts of the tree of life.

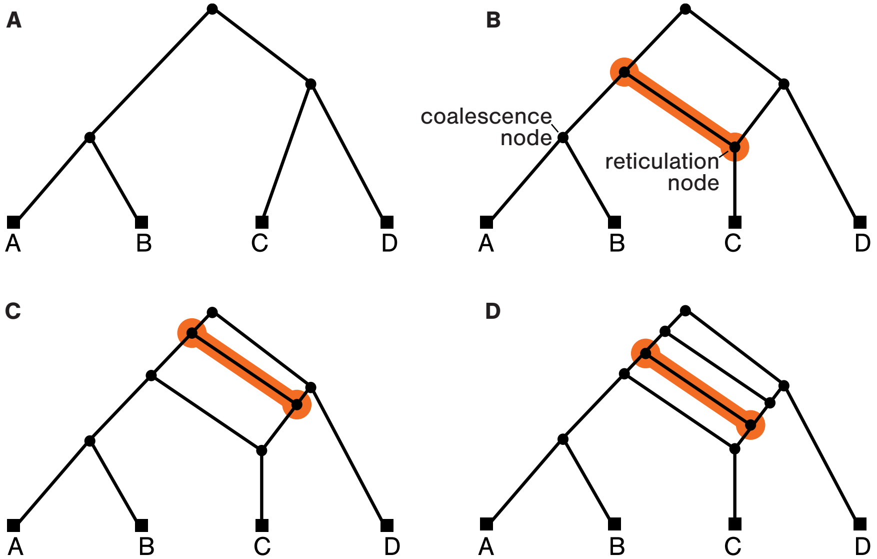


Figure 11.1: **A** A phylogenetic tree on four leaves can be transformed into a **B** phylogenetic network by adding a new edge and a reticulation node. **C, D** New network topologies can be created by adding new edges. This process can be continued endlessly, meaning there are infinite possible network topologies relating the four tips.

11.2.1 Homologous recombination processes

Homologous recombination in eukaryotes In eukaryotic organisms such as plants and animals, homologous recombination usually occurs due to genetic crossover during *meiosis*: the process that generates the germ cells necessary for reproduction. A schematic view of this process is shown in Figure 11.2.

A result of this process is that sites that are close to one another in a single chromosome sequence are more likely to be inherited together than sites that are far apart since distant sites are more likely to be separated by a crossover event. The propensity for sites to be inherited together is referred to as *linkage*. Nearby sites are, therefore, usually strongly linked, and the strength of this linkage decays with the distance between the sites. Sites on different chromosomes are almost completely *unlinked*, as the ancestry of a site on one chromosome is, after very few generations, almost completely distinct from sites on another chromosome.

Recombination in bacteria In bacteria, recombination is so prevalent that its influence on genetic diversity matches or outweighs that of mutation alone in many species (Vos and Didelot 2009). Homologous recombination is commonly used as a DNA repair mechanism

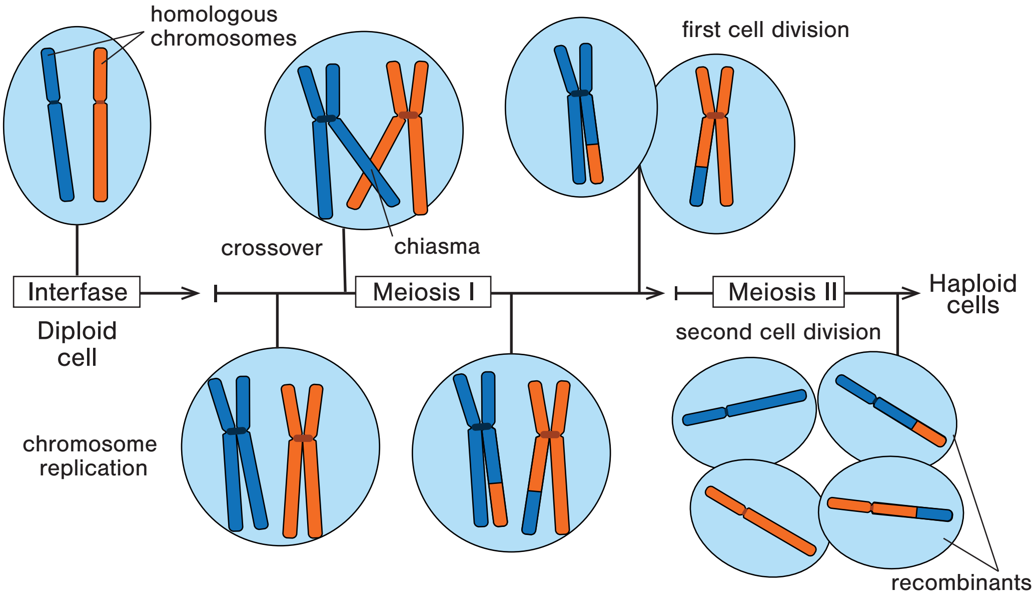


Figure 11.2: Homologous recombination. Schematic representation of the process of germ cell generation via meiosis. During this process, crossover results in the generation of new alleles by combining the homologous parental alleles.

and can also introduce variability into the bacterial genome. It occurs when a DNA molecule (such as the chromosome) is interrupted by a double- or single-stranded break, for instance, due to UV radiation. Other stretches of DNA in the cell that are homologous to the DNA on either side of the break will be used to patch the break.

If the used DNA originally came from a different parent cell, new genetic material is introduced into the DNA molecule. DNA from multiple parent cells can be present in the same bacterium due to horizontal gene transfer (HGT). Bacteria share genetic material between neighbours as a result of several distinct processes. These include:

Conjugation, in which pairs of bacteria exchange genetic material via cell-to-cell contact, mediated by a specially constructed tubular apparatus known as the pilus;

Transduction, in which genetic material is exchanged via bacteria-infecting viruses known as bacteriophages (or simply phages);

Transformation, in which a bacterium takes up naked genetic material from its surrounding environment.

Unlike crossover-style recombination in eukaryotes, homologous recombination in bacteria is often best described as *gene conversion*. This style of recombination results in replacing a fragment of the recipient's genome with a homologous fragment from the donor. This asymmetry is unlike the situation in eukaryotes, where material is swapped.

We note that HGT may lead to gene conversion, but not always. In many cases, the genes acquired through HGT are kept in addition to the genomic material of the recipient bacterium without recombining with the original genomic material. An example is the transfer of a whole plasmid (through conjugation), which then exists in the cytoplasm without gene conversion.

Recombination in viruses Recombination in viruses occurs often enough that it, too, must be considered when inferring phylogenies of virions. As discussed by Pérez-Losada et al. (2015), recombination occurs due to different mechanisms in different viruses but essentially results from single host cells becoming infected with multiple strains of the same virus, leading to the production of hybrid strains. Again, the recombination processes are subtly distinct from what occurs in eukaryotes and can generally be regarded as an example of gene conversion. Additionally, in segmented viruses, a particular mechanism, reassortment (see also below), may occur. This is not a recombination process as defined above, but it can be modelled in similar ways.

11.2.2 Impact of recombination on phylogenies

To develop an intuition for the potential effect homologous recombination has on phylogenetic inference, consider an alignment of just three eukaryotic sequences (Figure 11.3). As is often the case with nuclear genetic sequences obtained from eukaryotic cells, each sequence is much shorter than a full genome. Now consider starting from the left-most site in the alignment. Because a character at a single site on one chromosome can only ever be inherited from a single parent, the relationship between the observed characters at this first site on the three sequences can definitely be represented using a tree. Similarly, it is highly likely that the same tree will also have produced the characters belonging to the second site. However, at some point along the sequence, it is possible (or almost certain, depending on the length of the sequences) that somewhere in the ancestry, a lineage will take a different path due to a recombination break-point. This will have the effect that the tree associated with this and subsequent sites will differ.

Thus, recombination means that the phylogeny of the sequences is described not by a single tree but by a sequence of *local trees* associated with contiguous sequence stretches, as illustrated in Figure 11.3.

Several factors govern the probability of seeing more than one tree across an alignment. Shorter sequence alignments are less likely to be affected by recombination than longer alignments. Single nucleotide polymorphisms (SNPs) are usually assumed to be widely separated and thus are often considered completely unlinked — meaning that each SNP has a completely separate phylogeny (this is a necessary assumption for the validity of GWAS analyses

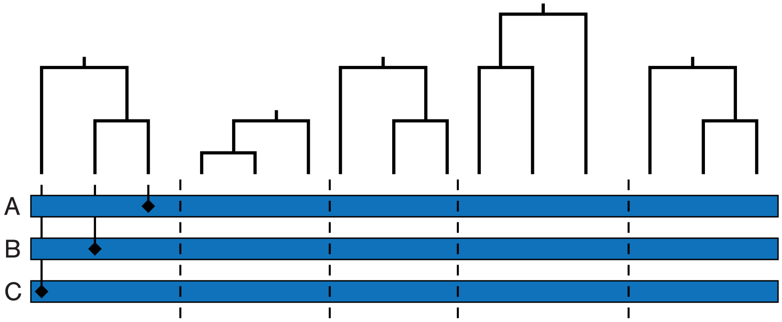


Figure 11.3: The effect of recombination on phylogenetic reconstruction. The blue bars represent an alignment of three sequences from individuals labelled A, B, and C. The phylogenetic relationships between the sequences are shown above the alignment. As a result of recombination, this relationship can be different in different alignment sections. Each of these trees is known as a local tree.

discussed in Chapter 4). Additionally, the rate of recombination in the study organism and the location of the sequence in the genome both play important roles.

When more than one local tree exists across a given alignment, recombination presents an important source of model misspecification, meaning that analyses that do not take this into consideration will inevitably produce biased results. Thus, phylogenetic analyses of recombining organisms must take steps to deal with these effects. Often, these steps involve using one of several pre-processing schemes to detect and remove parts of the sequence alignment affected by recombination.

One very well-known filtering scheme for bacterial recombination is implemented in Gubbins (`https://nickjcroucher.github.io/gubbins/`) (Croucher et al. 2014). It uses a sliding window approach to identify portions of the alignment that exceed some statistical incompatibility threshold with a maximum likelihood tree estimated using the full alignment. These portions are then removed, and a new maximum likelihood tree is estimated. This process is repeated iteratively until no incompatible portions remain.

While such schemes allow us to continue using our standard phylogenetic tree inference methods, they have the disadvantage of forcing us to discard potentially useful data. The alternative approach, which we will discuss in the remainder of this section, is to model the recombination process explicitly. This modelling is typically performed with the Wright-Fisher model and its limiting coalescent process. The birth-death model has not been considered widely in this context.

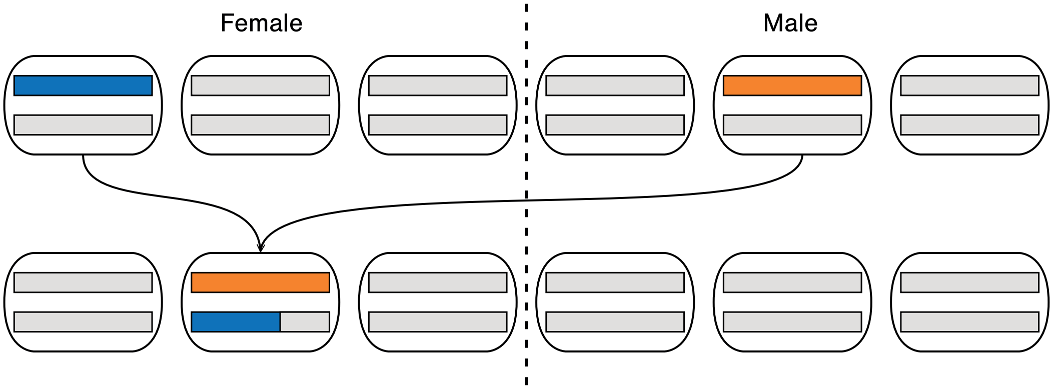


Figure 11.4: Explicit model of sexual reproduction and recombination. Boxes represent individuals, while the bars inside the boxes represent homologous pairs of genetic sequences.

11.2.3 Wright-Fisher model with recombination

We can model eukaryotic recombination using an extension (Hein, Schierup and Wiuf 2005) of the Wright-Fisher model introduced in Section 9.2. To introduce this extension, we first consider the diploid discrete-generation model illustrated in Figure 11.4. In this model, we explicitly divide the population into males and females, each individual carrying a homologous pair of genetic sequences. These sequences may be portions of the full genome. We aim to model the ancestral history of these genetic sequences.

We do this by selecting one member of the child generation and randomly assigning it one male and one female parent. Then, each of these chosen parents uses its own sequence pair to produce a single combined sequence. With probability r, which is proportional to the sequence length, this combined sequence contains a proportion of one of the homologues as the first part of the sequence, and the remainder is made up of the other homologue. The breakpoint is chosen uniformly at random. With probability $1 - r$, the combined sequence is made up entirely of the first or the second sequence (precisely which of the sequences is selected is determined uniformly at random). Finally, two combined sequences, one from each parent, are assigned to the child. The process is then repeated for all remaining children. This procedure is illustrated in Figure 11.4.

Just as with the original Wright-Fisher model, we now consider that the pairing of homologous sequences within specific individuals matters only over the course of a few generations and is largely irrelevant over evolutionary timescales. Thus, we can remove the boxes around pairs of homologous sequences and treat each independently without dramatically affecting the model, as shown in Figure 11.5. In this simplified model, we only consider the effect of the process on the individual sequences. Thus, each child sequence selects one parent sequence from the

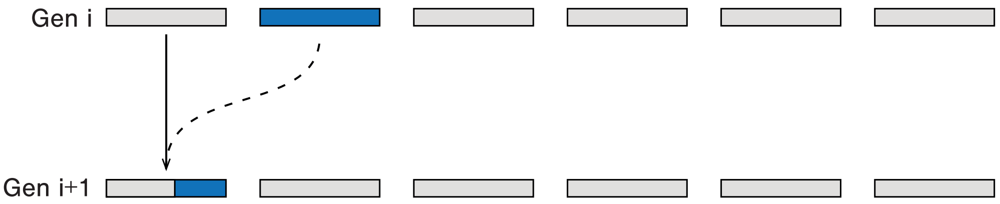


Figure 11.5: Wright-Fisher model with recombination. Individual sequences in the child generation randomly choose a parental sequence in the previous generation, then with probability r choose a second parental sequence to combine with the first.

previous generation. With probability r, the child selects an additional parent sequence and combines this with the first around a sequence breakpoint chosen uniformly at random. This simplification also allows the model to be applied to bacterial or viral populations, which cannot be divided between male and female parents.

We will refer to this simplified model as the Wright-Fisher model with recombination.

11.2.4 Coalescent with recombination

Just as we did in Section 9.2, we now consider the ancestry of a small number of individuals (or, more precisely, individual sequences) in the limit as the total number of individuals N becomes large and the time between generations g becomes small (with $\theta = {}^1\!/\!_{(Ng)}$ being constant). With the inclusion of recombination, we also require that the recombination probability r becomes small, such that the recombination rate $\phi = r/g$ remains finite.

In this limit, we arrive at the coalescent with recombination, first formulated by Hudson (1983). Like the regular coalescent, this is a backwards, continuous-time Markov process that begins with several sampled lineages and whose realisations describe possible ancestral relationships between those samples. However, unlike Kingman's coalescent (see Section 9.2.2), the coalescent with recombination produces recombination networks, also referred to as *ancestral recombination graphs (ARGs)*, instead of trees.

The coalescent with recombination involves two distinct stochastic events:

(i) coalescent events, which describe the point at which a pair of lineages finds a common ancestor, and

(ii) recombination events, which represent the effect of a recombination breakpoint within the sampled sequence.

For a given number of lineages k, coalescence occurs at the same rate as for the original coalescent process:

$$\binom{k}{2}\frac{1}{Ng}, \tag{11.1}$$

meaning that each pair of lineages coalesces with a fixed rate $^1/_{Ng}$. Recombination events, on the other hand, occur at the rate

$$\phi k, \tag{11.2}$$

meaning that each lineage splits (backwards in time) with a fixed rate ϕ. The process ends when the number of lineages reaches 1.

In addition to the merging and splitting of lineages, at a recombination event, the model must specify which sites are associated with each parental lineage. That is, it needs to specify which parent contributed the character at each site of the child sequence. This can be modelled by recording a sequence breakpoint with each (observed) recombination event and stating that the left-hand parent contributed everything to the left of this breakpoint, while the right-hand parent contributed everything to the right.

An example realisation of this process is shown in Figure 11.6. The bars below each edge represent the sequence carried by that edge. The blue intervals represent those portions of the sequence that are ancestral to at least one sampled sequence (all bars at the base of the graph are completely blue). Grey portions represent sites that are ancestral to no sampled sequences. The orange intervals represent sites that have found a common ancestor, meaning that the local tree belonging to these sites has already found a root. The root of the network is the *grand MRCA* or *GMRCA*: the first time that all ancestral lineages coalesce into a single individual. Note that the GMRCA can be much older than the roots of all local trees.

For further information on coalescent models for recombination networks, please refer to, for example, Hein, Schierup and Wiuf (2005).

11.2.5 Bayesian inference

Phylogenetic recombination networks can be inferred in the Bayesian framework using an approach very similar to the one described in Chapter 10. We begin by writing down the expression for the posterior distribution of a phylogenetic network G given a sequence alignment:

$$P(G,\phi,\theta,Q|A) = \frac{1}{P(A)}P(A|G,Q)P(G|\phi,\theta)P(\phi,\theta,Q), \tag{11.3}$$

where Q is the substitution rate matrix, ϕ is the recombination rate and θ the coalescence rate.

Each of the terms in the numerator of the right-hand side of this equation is easily evaluated.

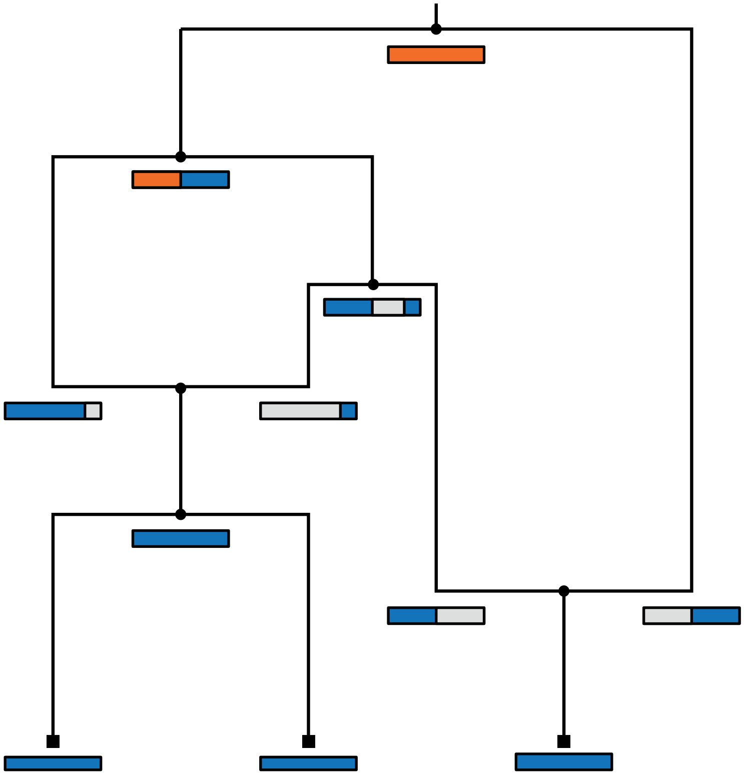


Figure 11.6: An example realisation of the coalescent with recombination. The bars below each edge represent the sequence carried by the lineage, with the blue interval representing those portions of the sequence ancestral to at least one sampled sequence. The orange intervals represent those sites that have found a common ancestor.

The network likelihood $P(A|G,Q)$ can be written as a product of local tree likelihoods:

$$P(A|G,Q) = \prod_{i=1}^{m} P(A_i|\mathcal{T}_i,Q), \tag{11.4}$$

where m is the number of local trees, $\mathcal{T}_i$ is the ith local tree, and A_i is the fragment of the sequence alignment corresponding to this tree. The probability density $P(G|\phi,\theta)$ can be evaluated analogously to the regular coalescent by considering the timings between coalescence and recombination events. Finally, the prior distribution $P(\phi,\theta,Q)$ has to be chosen as part of the model when setting up an analysis.

However, performing inference under this model is far more challenging than under the standard coalescent. This is because achieving chain convergence in a reasonable time is difficult due to the following reasons (among others):

(i) the space of phylogenetic networks with significant posterior probability is usually extremely large,

(ii) some features of G, including the exact times of recombination events, do not contribute directly to the likelihood (that is, some features are unidentifiable), and

(iii) the likelihood surface contains many distinct peaks.

Despite these challenges, several MCMC-based algorithms exist for inference under the coalescent with recombination, including:

- ARGWeaver (`https://github.com/mdrasmus/argweaver`) is an MCMC sampler under a computationally efficient approximation of the coalescent with recombination (Rasmussen et al. 2014);
- PSMC (`http://github.com/lh3/psmc`) is a method that also uses an approximation of the coalescent with recombination to perform phylodynamic inference of ancestral populations (Li and Durbin 2011);
- SMARTIE is an earlier MCMC algorithm that actually uses a non-informative network prior rather than the coalescent with recombination (Bloomquist and Suchard 2010);
- ClonalOrigin (`https://github.com/xavierdidelot/ClonalOrigin`) is an MCMC sampler for bacterial ARGs under the coalescent with gene conversion, which is a modification of the coalescent with recombination to account for homologous gene conversion — the kind of recombination most prevalent among bacteria (Didelot et al. 2010);
- CoalRe (`https://github.com/nicfel/CoalRe`) is an MCMC algorithm for inferring ARGs resulting from reassortment (rather than recombination) processes in viruses (Müller et al. 2020), where the underlying model is again based on the coalescent with recombination.

In the remainder of this section, we will discuss several concrete applications of some of these recombination network inference methods in different areas of biology.

11.2.6 Applications

11.2.6.1 Inference of ancestral human population dynamics (PSMC and MSMC)

One well-known application of the coalescent with recombination has been to infer human population dynamics using whole-genome sequence data. Recall that coalescent-based phylodynamic models can be used to infer ancestral effective population sizes, as discussed in Section 9.2. The signal for such an inference is derived from the timing between coalescence events drawn from exponential distributions whose rates are inversely proportional to the population size. In the absence of recombination, it would only ever be possible to obtain a very broad estimate of the effective population size from a single pair of aligned sequences since their ancestry would be described by a single two-tip tree with a single coalescence time.

In the presence of recombination, however, this situation changes dramatically. As shown in the schematic on Figure 11.3, recombination induces changes in the local tree over the alignment, meaning that, instead of a single coalescence event, we can have a large number of independent coalescence times between a single pair of aligned sequences given recombination occurs frequently. Since eukaryotes generally carry two distinct copies of every chromosome, this leads to the very tantalising possibility of inferring population dynamics from the genetic material of just one member of that population.

This precise idea was used by Li and Durbin (2011) to infer ancestral human population dynamics using homologous whole chromosome sequences assembled from blood samples from individual people. Their method, known as the *pairwise sequentially Markovian coalescent (PSMC)*, is based on an approximation to the coalescent with recombination but does not employ MCMC. Rather, it uses a technique known as *expectation maximisation* (Dempster, Laird and Rubin 1977) to find the population history function, which maximises the posterior probability distribution. Note that while this algorithm employs the coalescent with recombination model as a prior distribution over local trees, it implicitly averages over all possible ARGs instead of explicitly inferring this network.

The original method was limited to inference based on pairs of homologous sequences. However, since the original publication, a multi-sequence extension known as the *multiple sequentially Markovian coalescent (MSMC)* has been developed (Schiffels and Durbin 2014).

11.2.6.2 Recombination among *Escherichia coli* (ClonalOrigin)

The *coalescent with gene conversion* is a modification of the coalescent with recombination, accounting for homologous gene conversion (Didelot et al. 2010) in bacteria. Specifically, this model accounts for the fact that gene conversion results in the substitution of short homologous fragments from an external source. This is in contrast to the crossover-style recombination described by the coalescent with recombination model, where genetic material is swapped.

Application of the model revealed recent recombination-driven gene flow from a Shiga toxin-producing clade of *Escherichia coli* (+STEC) to a non-Shiga toxin-producing clade (-STEC) within BEAST2 Vaughan et al. (2017). This gene flow is potentially important because +STEC is a highly pathogenic form of *Escherichia coli*, and thus may result in currently non-pathogenic strains becoming pathogenic in the future. Figure 11.7 illustrates a summary network produced from the posterior distribution of this analysis.

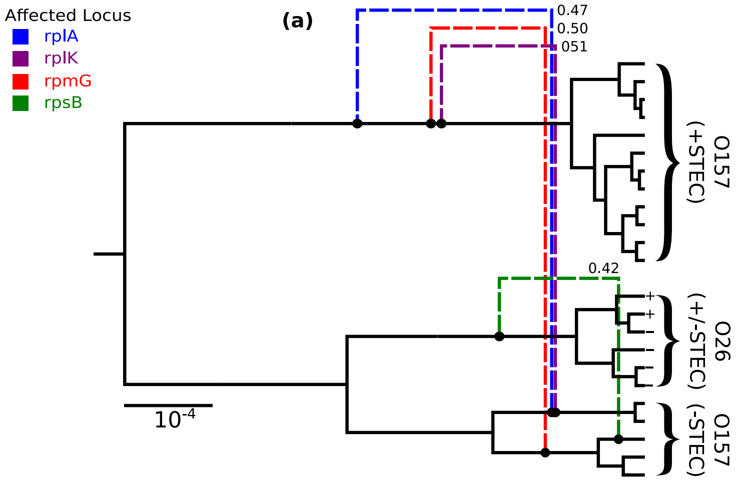


Figure 11.7: Inferred ancestral recombination graph representing the ancestry of *Escherichia coli* sequences. Edges due to recombination events are represented by dashed lines, with colours indicating the gene affected. The timing of the events indicates that gene flow is from +STEC to -STEC strains. Figure adapted from Vaughan et al. (2017).

11.2.6.3 Reassortment networks in influenza (CoalRe)

Next, we will consider *viral reassortment* (Lowen 2018), which is distinct from recombination. This process occurs in viruses whose genomes are divided into two or more distinct segments. This segmented structure allows cells that are infected by more than one viral strain to potentially produce hybrid strains whose genomes are composed of segments drawn from different original strains. Among the many viruses that display such segmentation are influenza A (8 segments) and Rotavirus (11 segments), both of which are very common in human populations (Varsani et al. 2018; McDonald et al. 2016).

Reassortment is interesting for two main reasons. Firstly, its presence means that one must take care when inferring phylogenetic trees from sequences drawn from segmented viruses since individual segments may have distinct phylogenetic ancestry. Secondly, reassortment is thought to play a key role in enabling zoonotic transmission, the transmission of viruses between host species.

Several methods exist for inferring reassortment networks from sequence data. Some approaches, such as TreeKnit (`https://github.com/PierreBarrat/TreeKnit.jl`) (Barrat-Charlaix, Vaughan and Neher 2022), take an indirect approach by inferring the network from trees inferred separately for each segment. This approach essentially looks for discrepancies between these segment trees and treats these as evidence for reassortment.

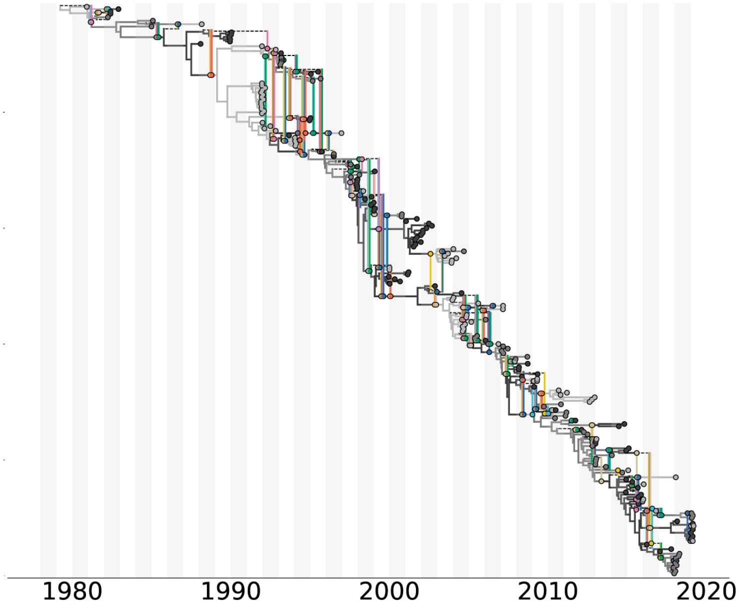


Figure 11.8: Estimate of the ancestral reassortment graph relating a set of influenza A subtype H3N2 genome samples collected over four decades, inferred using CoalRe. Figure adapted from Müller et al. (2020).

Müller et al. (2020) recently introduced a method named CoalRe (https://github.com/nicfel/CoalRe?tab=readme-ov-file) for inferring reassortment graphs directly from sequence data. This is an MCMC-based Bayesian approach very similar to that described in Section 11.2.5, with the only major difference being that CoalRe employs a modified version of the coalescent with recombination (Section 11.2.4) which accounts for the differences between eukaryotic recombination and viral reassortment. In particular, while the former randomly selects a breakpoint to divide the genome between the left and right parents, the latter assumes that each segment is randomly assigned to one or another of the parental lineages. This method has been extended by Stolz et al. (2021) to also account for population structure in the viral host population.

Figure 11.8 illustrates the application of this approach to the inference of the reassortment network ancestral to a set of sequences derived from isolates of influenza A subtype H3N2. This influenza subtype is one of the main influenza viruses responsible for seasonal influenza outbreaks. While most of the network looks like a regular phylogenetic tree, reassortment events (shown using coloured edges) are clearly visible. Note that, this being a Bayesian approach, the actual inference result is a posterior distribution over ancestral networks — the one shown in this figure is merely a summary network derived from this distribution.

11.3 Species networks

As mentioned at the start of this chapter, phylogenetic networks also exist at the level of species evolution. Some species populations can merge and form a hybrid species, which is very common in plants and some fish. This can occur, for instance, when populations that are close enough in terms of their genome and were previously geographically separated are brought into contact (this could happen through some geological process, such as a land bridge). Subsequent interbreeding can lead to a hybrid population, thus forming a new species from the genetic material of the two parent species. Additionally, asymmetric *horizontal gene transfer (HGT)* events may result in a number of genes from one species being transferred to another. The occurrence of these processes is one of the many reasons why species definition is difficult.

Before considering how to model these processes, first consider that even when representing species evolution using trees, one has to bear in mind that each particular species tree branch does not represent a single individual but, rather, a whole population of individuals of a species, and the branching events are not as obviously defined as when considering, for example, the division of individual bacteria. When tracing the history of sampled sequences of individuals from different species into the past, the coalescence of the genetic sequences may occur at a time earlier than the speciation time. Thus, we obtain two distinct types of trees, *gene trees* that follow the evolution of genetic sequences within species populations and *species trees* that follow the evolution of the species populations themselves (see Figure 11.9, ignoring, for now, the branch labelled with γ and the red gene tree). For more details on these nested trees, see, for example, Degnan and Rosenberg (2009).

If hybridisation or horizontal gene transfer between species occurs, the species ancestry needs to be modelled by a phylogenetic network instead of a tree, as trees alone cannot represent the merging of species. For instance, allowing for species hybridisation requires the species ancestry model to include special hybrid nodes that have two ancestral species and only one descendant species lineage — as shown in Figure 11.9.

Many methods for inferring species networks have been proposed over the years. For instance, Yu, Barnett and Nakhleh (2013) provided a method that sought to determine the most parsimonious network (fewest reticulations) given a set of gene trees. More recently, Bayesian methods capable of inferring species networks together with the nested gene trees have been introduced (see, for example, Wen, Yu and Nakhleh (2016), Zhang et al. (2018) and Rabier et al. (2021)). However, Bayesian approaches tend to be highly computationally demanding due to the extremely large state space of species networks with nested gene trees. This, therefore, remains an area of active development.

Despite these challenges, several groups have made exciting headway in using these methods to learn about species ancestry in the presence of hybridisation and horizontal gene transfer. A fascinating example can be found in the work of Barley et al. (2022), who apply (among many other approaches) the Bayesian PhyloNet (`https://phylogenomics.rice.edu/html/phylonetTutorial.html`) method by Wen, Yu and Nakhleh (2016) and

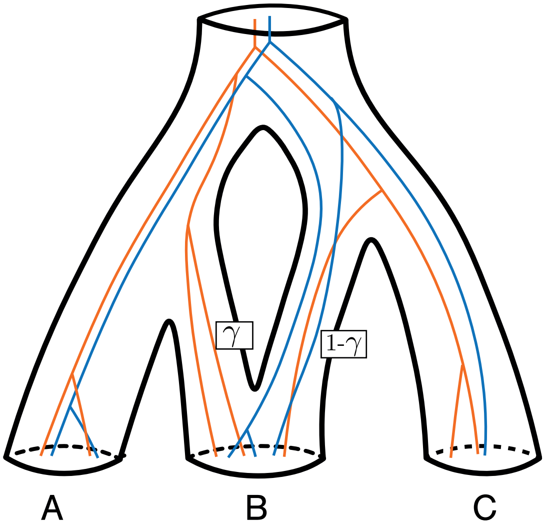


Figure 11.9: Species network (tubes) and its embedded gene trees (orange and blue lines), as used by Zhang et al. (2018) in their inference method. The parameter γ is the probability with which individual gene tree lineages are assigned to one of the parental species. Figure adapted from Zhang et al. (2018).

the likelihood-based PhyloNetworks (`https://crsl4.github.io/PhyloNetworks.jl/latest/`) method by Solís-Lemus and Ané (2016) to probe the reticulations present in the complex evolutionary history of over 30 species of North American whiptail lizards. Their Bayesian analysis provides evidence for a large number of distinct hybridisation events being responsible for the present-day diversity seen in these lizards. These and other examples demonstrate the importance of accounting for non-treelike processes when seeking to understand the evolution of species.

12 Opportunities and challenges

Throughout this book, the empirical examples of the methodology we have presented come mainly from the fields of macroevolution and virus epidemiology (in particular, we discussed RNA viruses such as HIV, HCV, influenza virus, Ebola virus, and SARS-CoV-2). Indeed, phylogenetic and phylodynamic approaches were first developed with a macroevolutionary application in mind. Later, these approaches were widely adopted and applied in RNA virus epidemiology. More recently, phylogenetic methods have been applied across various biological and non-biological fields. In the introduction, we briefly touched on some of these emerging application areas (Section 1.1.1). This last chapter will discuss these areas in more detail, linking to the concepts introduced throughout the book. As outlined below, phylogenetic approaches have been applied in these areas, yet the opportunities stemming from phylodynamic approaches have not been widely explored.

This chapter is comprised of two parts. The first highlights some of the many opportunities to achieve new understanding across biological and non-biological scales by employing the tools discussed in this book. The second part outlines major statistical and computational challenges in genetic sequence analysis with an evolutionary perspective in mind.

12.1 Opportunities through novel applications

12.1.1 Applications in biology and the life sciences

12.1.1.1 Infectious disease epidemiology beyond RNA viruses

Much of the epidemiological examples in this book involve RNA viruses. Indeed, the first epidemiological applications of phylogenetic and phylodynamic techniques involved RNA viruses, with DNA virus and bacterial applications appearing more recently. One essential reason for this is that RNA viruses evolve much faster than DNA viruses or bacteria. A single gene of RNA viruses typically contains enough diversity to perform phylodynamic studies (Chapter 9). Evolution is slower for DNA viruses and bacteria, and thus, diversity in samples is low. A single gene typically does not contain enough diversity to yield a phylogenetic signal on epidemic time scales; however, whole genomes may contain enough diversity. Thus, DNA virus and bacterial applications could only be pursued once whole genome sequencing became possible on a large scale.

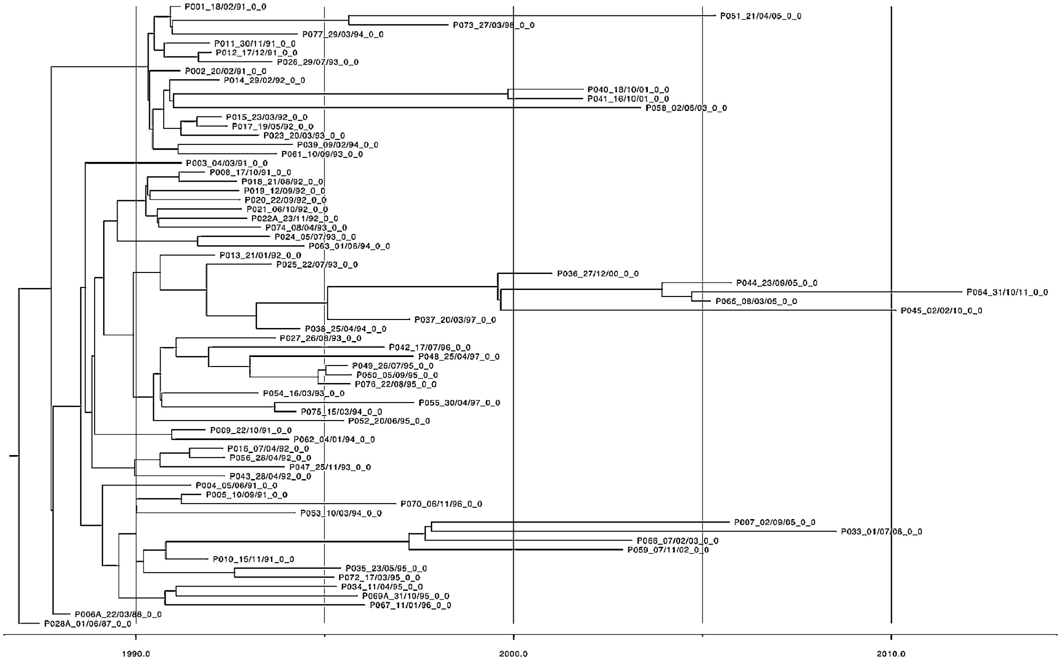


Figure 12.1: Phylogenetic tree of *Mycobacterium tuberculosis* strains collected during an outbreak in Switzerland. Inference was performed with Bayesian methodology. Figure adapted from Kühnert et al. (2018a).

The concepts introduced in this book can generally be used for the epidemiology of infectious diseases beyond RNA viruses, such as DNA viruses and bacteria, based on whole genome sequence data. As for RNA virus epidemiology, in this context, a unit is an infected host. For DNA viruses or bacteria without horizontal gene transfer (such as *Mycobacterium tuberculosis*), the presented phylogenetic and phylodynamic tools can be used directly (see Kühnert et al. (2018a) and Pečerska et al. (2021) for *Mycobacterium tuberculosis* examples). A calendar-time scaled phylogenetic tree of an *Mycobacterium tuberculosis* outbreak in Switzerland is shown in Figure 12.1. This study suggests that the peak of infections in this outbreak was around 1990, several years prior to the detection of the outbreak (as also suggested by Stucki et al. (2015)).

For bacteria with considerable horizontal gene transfer, the non-tree-like processes have to be acknowledged, as discussed in Chapter 11, where we also provided an application of phylogenetic networks to *Escherichia coli* (Figure 11.7).

Finally, eukaryotic infectious agents such as *Plasmodium* causing malaria cannot easily be analysed in the phylogenetic context: they recombine very frequently so that linkage between sites is very weak, and no tree or network represents the transmission chain well. However, GWAS approaches (Chapter 4) can be used to analyse such pathogen sequence data (Orjuela-Sánchez et al. 2010).

12.1.1.2 Microevolution

Microevolution refers to evolution within a single species population, with an individual member of the population as the biological unit. Any evolutionary study of bacterial populations falls under this application (e.g. Chapter 11). So does the study of evolutionary changes within the human population (again see Chapter 11 as well as Chapter 4) or a population of individuals from some other species. Furthermore, viruses within a host also form an evolving population.

Most bacteria exchange genetic material horizontally, most eukaryotes reproduce sexually, and viruses within a host may recombine when infecting the same cell. For pathogens, we were able to ignore such horizontal processes in most parts of the book because we focused on transmission trees where a tip corresponds to a host rather than a tip corresponding to a single pathogen individual. However, when studying, for example, an HIV population within a single host, where the individual is a single pathogen such as a virion, recombination cannot be ignored. Similarly, when considering the macroevolutionary scale with a species being an individual, we were considering species trees throughout the book (but see Chapter 11 for horizontal events in macroevolution); yet when performing microevolutionary studies focussing on individuals within species, horizontal processes are very prevalent. Network models (Chapter 11) are required in such microevolutionary settings unless only SNPs without linkage are considered, in which case GWAS approaches are valid Chapter 4.

12.1.1.3 Immunology

An immunological application of phylogenetic and phylodynamic methods is the proliferation of B lymphocytes (also called B cells), responsible for producing antibodies. Antibodies bind to antigens on the pathogen surface and neutralise them.

A wide repertoire of B cells is generated through so-called VDJ recombination in the bone marrow[1]. When a B cell is exposed to a pathogen, it undergoes somatic hypermutation. The resulting B cells ideally target the pathogen better (affinity maturation). This process of somatic hypermutation can be modelled by a tree: cells divide and rapidly accumulate point mutations. In this context, a B cell is a biological unit, and phylodynamics can be used to investigate the dynamics of B cell evolution during affinity maturation (Hoehn et al. 2016).

In Hoehn, Pybus and Kleinstein (2022), the authors infer B cell phylogenetic trees with maximum parsimony (see Figure 12.2). Parsimony is commonly employed in this context for speed reasons. However, maximum likelihood methods taking into account explicit models for B cells have been developed (Hoehn, Lunter and Pybus 2017; Hoehn et al. 2019), followed by

[1] Susumu Tonegawa received the Nobel Prize in Physiology or Medicine for this finding.

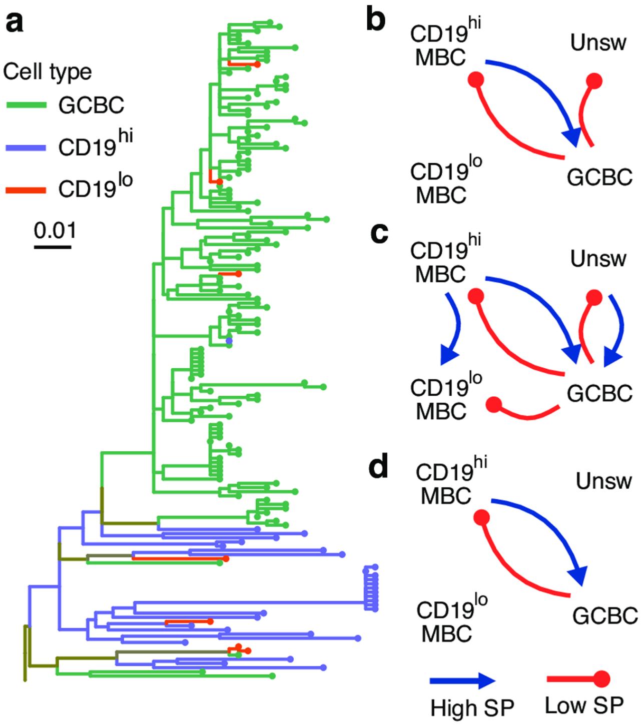


Figure 12.2: a Phylogenetic tree of B cells and **b**, **c**, **d** estimated ancestry of cell types in different patients. Figure adapted from Hoehn, Pybus and Kleinstein (2022).

Bayesian methods (Dhar et al. 2020). Alongside these methods, simulation frameworks to assess the performance of the new approaches for B cells were developed, such as Yermanos et al. (2017). Advances in sequencing B cells and robust inference of their phylogenies offer unique opportunities to enhance the understanding of affinity maturation by employing phylodynamic approaches (Stadler, Pybus and Stumpf 2021).

12.1.1.4 Developmental biology

In the development of an organism, an initial cell (the fertilised egg) develops through cell division, differentiation, and death into a full multicellular organism. Thus, the ancestry of all cells within an organism can be depicted in a tree with single cells as the unit represented as a tree branch. For *Caenorhabditis elegans*, the whole developmental tree has been painstakingly mapped by directly imaging its developmental process (Sulston et al. 1983). However, this is a unique exception: such imaging is generally not possible over the whole developmental time. An alternative is to reconstruct the tree using genomic sequences; however, few somatic mutations occur during healthy development. Thus, it is very challenging to reconstruct a cell tree based on the genome sequences of different cells from an organism.

Recent technological advances aim to generate synthetic sequence data that are informative enough to reconstruct the tree of development. The underlying principle is to insert an engineered "barcode" into the genome of the first cell. This barcode is constructed such that it evolves in a neutral way and accumulates enough diversity to reconstruct the cell phylogenetic tree based on sequenced barcodes from different cells (McKenna and Gagnon 2019; McKenna et al. 2016; Raj et al. 2018; Spanjaard et al. 2018; Alemany et al. 2018; Chow et al. 2021; Loveless et al. 2021; Choi et al. 2022). These barcodes, in principle, contain information about the cell tree of a whole organism or the cell trees of particular organs or other tissue.

The statistical tools for molecular evolution presented in this book (Chapter 5) need to be adapted to analyse the barcode genomic data since barcode evolution is typically not dominated by point mutations but instead by processes such as scarring, insertion, and deletion. Mostly parsimony or distance-based methods have been used on the barcode datasets (e.g. Choi et al. (2022), also shown in Figure 12.3). Recently, model-based approaches such as Feng et al. (2021) were developed, some also providing a calendar-time scale for the tree (Seidel and Stadler 2022; Fang et al. 2022).

A core challenge in developmental biology is to understand and quantify cell differentiation. Datasets containing both the barcodes and the transcriptome (from RNA sequencing) for a set of cells allow us to reconstruct calendar-time scaled phylogenies based on the barcodes; the trees can be amended with the transcriptomic information for the tips. Such a tree provides insights into how cells differentiated into their current transcriptomic state (Kester and van Oudenaarden 2018). Using the new barcoding and sequencing technologies together with phylodynamic models will open up the opportunity to obtain a quantitative understanding of the cell differentiation process. A phylodynamic cell differentiation model, the *cell state transition diagram*, based on the compartmental models in Section 9.5.2, was introduced in Stadler, Pybus and Stumpf (2021).

12.1.1.5 Cancer

Cancer occurs as a result of an out-of-control cell division process. During this uncontrolled division, many genetic changes occur, such as point mutations, copy number variations, chromosomal rearrangements, and ploidy changes. The genomes of cancer cells can be sequenced to reconstruct the underlying cell phylogenetic tree, where single cells are the unit corresponding to a branch in the tree.

Classically, bulk sequencing of cancer samples was performed. Bulk sequencing means that based on a sample, raw reads are obtained from a set of cells jointly. These reads then represent the cells within the considered sample. Treating each sample as a tip does not lead to the evolutionary tree since samples contain information on a set of cells that may not be clonal;

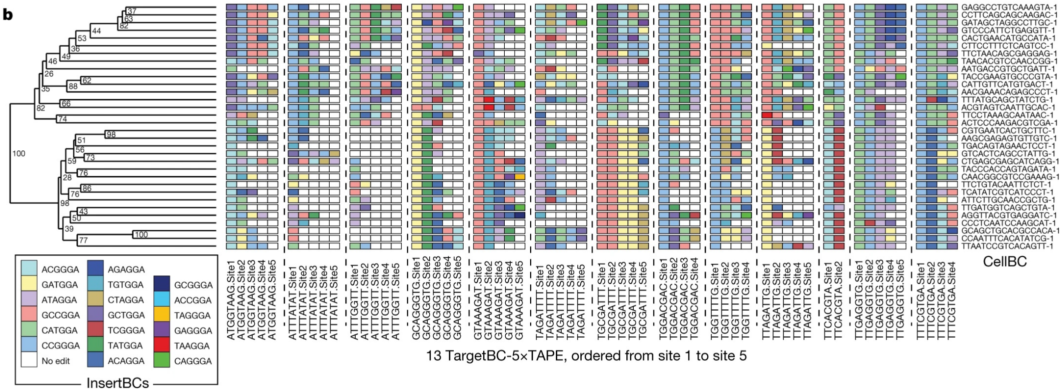


Figure 12.3: Phylogenetic tree of single cells traced with a barcode during the developmental process. Figure adapted from Choi et al. (2022).

instead, some cells in one sample may be more closely related to cells in another sample compared to cells in the same sample (Alves, Prieto and Posada 2017). Computational methods were developed, aiming to deconvolve the mixture such that reads are assigned to different cells (Sottoriva et al. 2015; Ling et al. 2015; Zhai et al. 2017; Beerenwinkel et al. 2014).

Tree structures, where essentially each node corresponds to a cell, were reconstructed from this deconvolved bulk data using different clustering approaches. The field is moving towards inferring the phylogenetic tree representing evolutionary history through time, now with cells being the tips, as done in Martinez et al. (2018) (based on so-called crypt-based bulk data) using a parsimony approach as well as a Bayesian approach within BEAST (see Chapter 10). For an overview, see Schwartz and Schäffer (2017).

Recent technology enables single-cell sequencing, resulting directly in one assembled sequence per cell based on which a tree can be reconstructed; see, for example, Casasent et al. (2018). A challenge is the noise in the single-cell sequence data due to the small amount of DNA. Computational methods such as Zafar et al. (2017), Kozlov et al. (2022), Kang et al. (2022) and Kang et al. (2023) introduce approaches to model this noise directly — meaning the downstream results take the noise into account — while also adapting substitution models to the specific evolutionary process in cancer. The latter is a prerequisite for accurate estimation of time-scaled trees.

Going forward, we see a lot of potential in analysing time-scaled cancer tumour trees from different patients jointly to quantify general patterns about cancer evolution and progression; see also Stadler, Pybus and Stumpf (2021).

12.1.2 Applications in anthropology

Concepts introduced in this book have also been used in anthropological studies. Here, we will outline advances in linguistics and human migration. Phylodynamics has further been used to study the evolution of stratified societies (Watts et al. 2016) and the evolution of political systems (Currie et al. 2010). Moreover, phylodynamics has been used to investigate the impact of past human trade on ocean ecology (for example, on herring species, (Atmore et al. 2022)). Finally, in a twist of irony, phylogenetics has been used to study the recent evolution of policies opposed to teaching evolution in schools (Matzke 2016).

12.1.2.1 Linguistics

Linguists encode languages into binary sequences, where each site on the sequence indicates the presence or absence of a sound/meaning combination known as a cognate in each language. Instead of a biological unit, we define a language as the anthropological unit we study. Substitution processes similar to the ones used to model nucleotide substitution are used to model the change in a sequence representing a language, representing the gain and loss of cognates from languages over time (Forster and Renfrew 2006; Nicholls and Gray 2008; Bouckaert and Robbeets 2017).

Based on an alignment of the sequences corresponding to languages with a substitution model and a tree-generating model, phylogenetic and phylodynamic inference can be performed as outlined in Chapters 6, 9 and 10.

A well-known example of this kind of application can be found in the study of the origin of the Indo-European language family (Bouckaert et al. (2012, Figure 2)).

12.1.2.2 Human migration

The past movement of humans can be estimated indirectly via the evolution of the microorganisms — specifically bacteria — which they carry. This approach has the advantage that the evolutionary processes affecting bacterial genomes generally occur thousands of times faster than the analogous process affecting human genomes. This means that evolutionary relationships between bacteria are much easier to resolve over the relatively short timescales involved in human migration than the corresponding human evolutionary relationships would be. A phylogenetic study of bacteria associated with human populations describes the movement of bacteria around the world and, thus, indirectly, human migration. In this context, a unit is a human population. In Linz et al. (2007, Figure S3), the past human migration process is reconstructed from *Helicobacter pylori* sequences using the neighbour-joining tree reconstruction method.

12.2 Statistical and computational challenges

The statistical advances across phylogenetics and phylodynamics presented in this book have been implemented as free software. Over the last decade, much of this work has focused on developing and contributing to large extensible software platforms. Major examples include RevBayes (https://revbayes.github.io/) (Höhna et al. 2016), MrBayes (https://nbisweden.github.io/MrBayes/) (Huelsenbeck and Ronquist 2001), BEAST1 (https://beast.community/) (Drummond and Rambaut 2007) and BEAST2 (https://www.beast2.org/) (Bouckaert et al. 2019) (the authors of this book mainly contribute to the latter), with all four being Bayesian frameworks. For the prospective BEAST2 user, we offer a "Taming the BEAST" website, housing many BEAST2 tutorials that can be used for self-study and are taught in the corresponding workshop series of the same name (Barido-Sottani et al. 2018). While such platforms were extended by many different researchers in various ways, offering broad functionality, core statistical and computational challenges in phylogenetics and phylodynamics remain.

12.2.1 Analysing large datasets

The Bayesian approach is very appealing due to its conceptual simplicity and its requirement to explicitly state the stochastic model under which the data are assumed to be generated, including the prior information. However, many implementations of Bayesian approaches quickly become unwieldy when applied to large datasets. As discussed in Section 6.2.3.3, the number of trees on n leaves grows exponentially on the order of $\mathcal{O}(e^{n \ln n})$. Traditional Metropolis-Hastings MCMC approaches, in particular, may require extremely long chains to characterise posterior distributions over such big tree spaces fully.

For this reason, statisticians have developed sophisticated alternatives. One example is the so-called Hamiltonian Monte Carlo approaches (Girolami and Calderhead 2011), which use the gradient of the target distribution to guide proposals and more efficiently explore continuous state space. However, despite advances in applying such methods to discrete phylogenetics state spaces (Dinh et al. 2017), practical Bayesian phylogenetic inference remains limited to trees with less than a thousand tips.

Distance-based methods and approximate maximum likelihood methods are much faster in estimating a phylogenetic tree topology. In fact, significant progress was made in estimating large tree topologies during the COVID-19 pandemic, as millions of SARS-CoV-2 sequences became available. One approach to continuously update very large trees with new sequences was to sequentially add new sequences to the tree using a parsimonious concept (https://genome.ucsc.edu/cgi-bin/hgPhyloPlace (UShER) (Turakhia et al. 2021)).

For datasets with a lot of signal and little uncertainty regarding the tree topology, a way forward can be to fix the topology using one of these fast methods and only account for uncertainty in branch lengths, evolutionary parameters, and phylodynamic parameters with the Bayesian approaches.

12.2.2 Combining genetic sequences with additional data

Genetic sequences encode aspects of the history of the individuals they represent. However, there is also a limit on the amount of information that genetic sequences contain; for example, as we discussed in Section 9.1.5.2, only two out of the three phylodynamic parameters (birth, death, and sampling) can be inferred from genetic sequence data alone. Additional data, such as fossil data or classic epidemiological data, can provide information on the third parameter.

Generally, many aspects beyond the genetic sequences may be measured to understand the evolutionary and population dynamic processes giving rise to the considered samples. For example, human travel may be considered when studying epidemics, RNA transcriptomic data may be generated for single cells, and climate or tectonic activity may be estimated for the earth over the past millions of years.

Conceptually, there are two different types of additional data. Each sequence may have an array of metadata associated with it (e.g. RNA transcriptomic data for single cells, clinical information on infection outcome or the age of infected individuals for epidemics, phenotypic traits for species). Alternatively, general data may be available spanning the time scale of the phylogeny (e.g. paleoclimatic data for macroevolution, prevalence information or travel patterns for epidemiology).

12.2.2.1 Additional data spanning the time scale of the phylogeny

For data spanning the time scale of the phylogeny, one can employ time-dependent phylodynamic models, where changes in dynamics through time are informed by the additional data. In a Bayesian context, the additional information flows into the phylodynamic model through the prior on time-dependent rates.

A factor that complicates analyses with such metadata comes into play when the metadata observations are not independent of the sequence data observations. Observing the climate through time is, of course, a measurement that is independent of sequence data of species. However, if we have case count data in addition to the pathogen sequences, then the case counts and the sequences are part of the same transmission tree and, thus, are not independent observations. If the dependency is weak (e.g. the number of sequences is very small compared to the number of counted cases), one can assume independence (Rasmussen, Ratmann and Koelle 2011) and work with time-dependent phylodynamic models. More generally, we need to acknowledge the dependence, which brings statistical and computational challenges (Vaughan et al. 2019).

12.2.2.2 Additional data for each tip in the phylogeny

One can use structured models (Section 9.5) for sequence-specific metadata, assuming a separate deme for each unique metadata combination. However, this quickly leads to an explosion in the number of parameters that need to be inferred.

The general approach to dealing with this is to additionally incorporate some prior knowledge about the similarity of the dynamics of individuals having similar (yet not identical) metadata combinations.

Assigning individuals to a discrete space (finite number of demes) A commonly used approach is to partition the metadata space into a small number of subsets. The approach then assumes that all individuals belonging to one of these subsets possess identical dynamics. Each subset is identified as a "deme" for the phylodynamic analysis. Multi-type analyses can then be feasibly performed on these demes, using either a structured coalescent or a multi-type birth-death model, as discussed in Section 9.5.

While this approach is straightforward, the obvious limitation is its requirement for partitioning the metadata. In some cases, there is a natural way to divide the metadata. For instance, consider the case where the metadata are geographical coordinates (latitude and longitude) of an animal population spread across several islands. Grouping all coordinates corresponding to the same island might make sense if the within-island populations themselves are homogeneous. However, imagine now that the geographical coordinates belong to a population spread across a large continent. In this case, a natural way to partition the coordinates, such that the assumption of homogeneous mixing within a subset makes sense biologically, may not exist.

As a side note, continuous phylogeography models such as those mentioned in Chapter 9 implicitly assign the entire geographic coordinate space to a single deme. Thus, these models assume the same phylodynamic parameters (such as birth, death, sampling or pairwise coalescence rate, and migration/diffusion rates) for all individuals and merely quantify the geographic movement of the individuals through time.

Assuming tight priors in continuous space (e.g. generalised linear models) A second approach assumes that each individual with unique metadata has its own parameters, meaning each individual exists in a separate deme. Then, tightly informative priors on a large number of phylodynamic parameters are assumed. This can be done by introducing some simple, functional relationship between the metadata combination and the phylodynamic parameters. For example, let us assume that metadata for each sequence is its geographical coordinates with the average temperature and yearly precipitation and that we have a simple model for how individuals move geographically. One may then assume that a linear function of the average temperature and yearly precipitation at a coordinate determines the (logarithm of the) transmission rate at that coordinate. This kind of approach is known as a generalised linear model (GLM, Nelder and Wedderburn (1972)), and it has been used in Lemey et al. (2014) to explore, for example, the relationship between migration history of the virus and factors such as airline flight frequencies, population sizes, and so on within a phylogeography approach (see Section 9.5.4.1). In this framework, regularisation techniques can be used to restrict the parameter space further, meaning non-zero parameters are penalised.

When employed in a birth-death model framework, the metadata can be used as information on the birth and death rates, meaning they may influence the birth and death rates. Thus far, we have considered the sequence alignment separately, as data evolving on a given tree (see models in Chapter 5), not influencing birth and death rates. In other words, it was assumed that the sequences evolve neutrally — the genotype influences neither birth nor death. However, one may treat parts of the sequences as part of the "metadata" and thus allow it to affect the phylodynamic parameters, lifting the assumption of complete neutrality. For example, in Section 9.5, we put sequences into different demes depending on whether they carry a drug-resistant mutation or not. This idea has been recently generalised by looking at several mutations simultaneously to infer genotype-dependent transmission rates in the context of Ebola and influenza virus evolution (Rasmussen and Stadler 2019).

The approach of assuming tight priors is still not widely used. A reason for this is the specialised nature of such models: they generally involve many parameters which must be tuned carefully for a particular application. Generally flexible and performant implementations of these methods are not yet available.

12.2.3 Determining model adequacy

As stated above, Bayesian inference yields important advantages over other inference frameworks due to transparency with respect to its modelling and prior assumptions and its explicit accounting for uncertainty in the final results. It also provides a clear framework, through the Bayes factors (see Section 10.1.6), for selecting from different possible models. But what happens when the models do not describe the biological system well? How much can we really trust our inference results?

One approach to this problem involves assessing what has been termed *model adequacy* (Bollback 2002).

This involves quantifying the degree to which the best-fitting model actually produced the observed data. It is centred around what is known as the *posterior predictive distribution* defined as the probability distribution of (hypothetical) new data conditional on the model and the existing data. Assessing model adequacy generally involves sampling the posterior predictive distribution and comparing key summary statistics of this new (simulated) data with the summary statistics of the observed data. The assumption is that if the model really does describe the observed data well, these summary statistics should be comparable. Practically speaking, sampling the posterior predictive distribution can be done by taking combinations of model parameters sampled from the posterior and simulating a new dataset under the model from each combination.

Several applications of this idea in a phylogenetic and phylodynamic context have been recently explored, and many more are in development. A technique for assessing the model adequacy of molecular clock models has been developed (Duchêne et al. 2015), as has a related technique applied specifically to birth-death (Duchêne et al. 2018), coalescent skyline plot (Fonseca et al. 2022) and phylogeographic (Carstens et al. 2022) models. While these tools are not yet widely used, such approaches will be instrumental in evaluating the appropriateness of models and, in turn, avoiding biased results due to model misspecification.

12.3 Final words

In recent years, the field of genomic sequence analysis has flourished thanks to new sequencing technologies producing an unprecedented wealth of data and exciting developments in statistics and computation. In this book, we have outlined classic approaches and recent developments that allow us to learn about biology by looking at the genetic code of different organisms from an evolutionary perspective.

We envision that the concepts in this book, combined with overcoming challenges outlined in this last chapter, will enable statistical and computational analyses of genomic sequence data across biological scales. More broadly, the tools will enable us to answer questions that lie beyond biology, studying further fields where evolution plays a role, such as anthropology. Overall, leveraging new datasets and statistical tools will enable us to answer fundamental research questions as well as generate evidence for policymakers.

List of Figures

1.1 Ape phylogeny 2
1.2 Guide through the book 6
1.3 Darwin’s evolutionary tree sketch
🅮 This figure is in the public domain, adapted from
https://commons.wikimedia.org/wiki/File:Darwin_Tree_1837.png 8
1.4 Darwin’s sketch of a time tree.
🅮 This figure is in the public domain, adapted from
https://commons.wikimedia.org/wiki/File:Arvore_darwin.jpg 9
1.5 Lamarckian evolution
🅭🅯🄎 This figure is distributed under the terms of the CC-BY-SA licence
(https://creativecommons.org/licenses/by-sa/4.0/), adapted from
https://commons.wikimedia.org/wiki/File:Dibujo_explicativo.jpg 11
1.6 DNA double helix 14
1.7 Codon sun 15
1.8 Central dogma of biology 16

2.1 Polymerase chain reaction (PCR) 30
2.2 Sanger sequencing 33
2.3 Pyrosequencing 36
2.4 SOLiD sequencing 38
2.5 Ion Torrent sequencing 39
2.6 Illumina sequencing 40
2.7 PacBio sequencing 42
2.8 Nanopore sequencing 43

3.1 Pairwise alignment example 47
3.2 Schematic of global and local alignment 47
3.3 Dot-matrices visualising similarity of two sequences 49
3.4 Needleman-Wunsch algorithm: score matrix example 53
3.5 Needleman-Wunsch algorithm: alignment reconstruction example 54
3.6 Smith-Waterman algorithm: score matrix example 56
3.7 BLOSUM62 matrix 59
3.8 BLAST: Example of the word extending step 60

4.1 Manhattan plot of p-values from a GWAS 73

5.1 Markov chain model for molecular evolution . . . 99
5.2 Substitution rates of the JC69 model . . . 100
5.3 Example of transition probabilities under the JC69 . . . 101
5.4 Example of transition probability matrices under the JC69 . . . 102
5.5 Substitution rates of the K80 model . . . 103
5.6 Illustration of time reversibility . . . 107
5.7 Substitutions along a phylogenetic tree . . . 113
5.8 Example alignment of two sequences and their ancestor . . . 115
5.9 Maximum likelihood estimation of the distance between two sequences . . . 118
5.10 Adaptation of the HIV envelope protein to the immune system response 🅭🅯 This figure is distributed under the terms of the CC-BY licence (https://creativecommons.org/licenses/by/4.0/), adapted from Wibmer et al. (2013) . . . 123
5.11 All codons with a single nucleotide difference to codon CGG . . . 127
5.12 All codons with a single nucleotide difference to the example codons . . . 131

6.1 Primate phylogeny . . . 136
6.2 Phylogenetic tree of HIV samples from different patients . . . 137
6.3 Phylogenetic tree of SARS-CoV-2 genomes © This figure is copyrighted to the authors of the article and used with permission, adapted from Hodcroft et al. (2021) . . . 138
6.4 Unrooted phylogenetic tree on four tips and a corresponding rooted tree . . . 139
6.5 The unrooted trees with two, three, and four tips . . . 141
6.6 Unrooted four-tip trees obtained from unrooted three-tip tree . . . 142
6.7 UPGMA example: intermediate tree after one iteration . . . 149
6.8 UPGMA example: intermediate tree after two iterations . . . 150
6.9 UPGMA example: final tree . . . 151
6.10 Parsimony example: tree obtained by the UPGMA algorithm . . . 156
6.11 Parsimony example: minimal number of substitutions for the UPGMA tree . . . 156
6.12 Parsimony score does not depend on rooting . . . 157
6.13 The three unrooted labelled trees on four tips . . . 157
6.14 Parsimony scores for the three unrooted trees on four tips . . . 162
6.15 Long branch attraction . . . 163
6.16 Likelihood example: unrooted tree used in the calculations . . . 165
6.17 Likelihood example: adding internal sequences and a root . . . 166
6.18 Setup of Felsenstein's pruning algorithm . . . 169
6.19 Evaluating probabilities with Felsenstein's pruning algorithm . . . 169
6.20 Time reversibility in the context of likelihood calculations . . . 171
6.21 Non-identifiability of clock rate and time . . . 173
6.22 Estimation of the clock rate for serially sampled sequences . . . 174
6.23 Nearest-neighbour interchange (NNI) move . . . 176
6.24 Subtree pruning and regrafting and tree bisection and reconnection move . . . 177

6.25 HIV/SIV phylogenetic tree
🅭🅯🄎 This figure is distributed under the terms of the CC-BY-SA licence (https://creativecommons.org/licenses/by-sa/4.0/), adapted from https://commons.wikimedia.org/wiki/File:HIV-SIV-phylogenetic-tree_straight.svg 180

7.1 Likelihood ratio test: example with a fair die 188
7.2 Likelihood ratio tests to identify an appropriate substitution model 189
7.3 Confidence interval estimates: example with a fair die 194
7.4 Bootstrap for phylogenies: concept 197
7.5 Bootstrap for phylogenies: example
© This figure is copyrighted to the authors of the work and used with permission, adapted from Sägesser (2010) 199

8.1 Assessing discrete character associations: example with eye and feather colours 202
8.2 Assessing discrete character associations: taking into account the phylogeny . 204
8.3 Assessing continuous character associations: wing span and beak length . . . 207
8.4 Brownian motion along a phylogeny 210
8.5 Contrast method: example tree 211
8.6 Contrast method: the pruning step 216
8.7 Contrasts for the example of wing span and beak length 218

9.1 Constant rate birth-death model 225
9.2 Population dynamics and trees generated by the birth-death model 226
9.3 Partitioning of time as employed in derivation of $p(0|t)$ and $p(1|t)$ 228
9.4 Partitioning of time as employed in the derivation of $p(n|t)$ 230
9.5 Trees generated under the birth-death phylodynamic model 233
9.6 Ranked labelled trees with four tips 235
9.7 Trees and their lineages-through-time plot 237
9.8 Lineages-through-time plot 238
9.9 Estimating birth and death rates from an LTT plot 244
9.10 Phylodynamic likelihood calculation: oriented tree example 245
9.11 Partitioning of time employed in the derivation of $p(t, x_1)$ 246
9.12 Phylodynamic estimation of the basic reproductive number in real-time . . . 253
9.13 Phylodynamic estimation of the basic reproductive number
🅭🅯 This figure is distributed under the terms of the CC-BY licence (https://creativecommons.org/licenses/by/4.0/), adapted from Vaughan et al. (2024) 254
9.14 Phylodynamic estimation of speciation and extinction rates
© This figure is copyrighted to the authors of the article and used with permission, adapted from Stadler (2011) 255
9.15 Wright-Fisher model: concept 257
9.16 Wright-Fisher model: an example 258
9.17 Coalescent calculations: example tree 261

9.18 Trees generated by coalescent processes assuming different population dynamics 264
9.19 Population dynamics using the skyline plot 265
9.20 Birth-death process trajectory and coalescence times 268
9.21 Coalescent skyline plot example: infected population size for HCV in Egypt . 270
9.22 Birth-death skyline plot example: effective reproductive number for HCV in Egypt 271
9.23 Population structure impacting the shape of a phylogeny 275
9.24 Transmission trees simulated under transmitted as well as *de novo* drug resistance 276
9.25 Reconstructed transmission trees for drug-sensitive and drug-resistant samples277
9.26 Multi-type birth-death model 278
9.27 Multi-type birth-death probability density calculations: example tree 281
9.28 Effective reproductive number of seasonal influenza
© This figure is copyrighted to the authors of the article and used with permission, adapted from Kühnert et al. (2016) 285
9.29 Phylogenetic tree of seasonal influenza
© This figure is copyrighted to the authors of the article and used with permission, adapted from Kühnert et al. (2016) 286
9.30 Structured Wright-Fisher model 287
9.31 Structured coalescent probability density calculations: example tree 291
9.32 Structured coalescent model analysis of influenza sequences
🅭🅯 This figure is distributed under the terms of the CC-BY licence
(`https://creativecommons.org/licenses/by/4.0/`), adapted from
Vaughan et al. (2014) 295

10.1 Prior and posterior distribution for genetic distance between two sequences . 305
10.2 Posterior distribution and posterior density intervals 308
10.3 Posterior distribution and normalisation factor 312
10.4 Proposals in an MCMC analysis 314
10.5 State trajectories and histograms of MCMC analyses 316
10.6 State trajectories and histograms of MCMC analyses with different proposals 318
10.7 Posterior distribution of H1N1 influenza A phylogenetic tree 321
10.8 Bayesian analysis of Ebola sequences from West Africa outbreak 2013-2016
© This figure is copyrighted to the authors of the work and used with permission, adapted from du Plessis (2016) 323
10.9 Bayesian analysis of SARS-CoV-2 sequences from Switzerland 2020
🅭🅯 This figure is distributed under the terms of the CC-BY licence
(`https://creativecommons.org/licenses/by/4.0/`), adapted from
Nadeau et al. (2023) 324
10.10 Penguin phylogeny obtained with the fossilised birth-death model
© This figure is copyrighted to the authors of the article and used with permission, adapted from Gavryushkina et al. (2017) 325

11.1 Phylogenetic networks on four leaves 329
11.2 Homologous recombination 330
11.3 Effect of recombination on phylogenetic reconstruction 332
11.4 Explicit model of sexual reproduction and recombination 333
11.5 Wright-Fisher model with recombination 334
11.6 Coalescent with recombination 336
11.7 Ancestral recombination graph obtained from *Escherichia coli* sequences
🅭🅯 This figure is distributed under the terms of the CC-BY licence
(`https://creativecommons.org/licenses/by/4.0/`), adapted from
Vaughan et al. (2017) 339
11.8 Ancestral reassortment graph of influenza A subtype H3N2 sequences
🅭🅯 This figure is distributed under the terms of the CC-BY licence
(`https://creativecommons.org/licenses/by/4.0/`), adapted from
Müller et al. (2020) 340
11.9 Species network with embedded gene trees
© This figure is copyrighted to the authors of the article and used with permission, adapted from Zhang et al. (2018) 342

12.1 Phylogenetic tree of *Mycobacterium tuberculosis* strains from an outbreak in Switzerland
© This figure is copyrighted to the authors of the article and used with permission, adapted from Kühnert et al. (2018a) 346
12.2 Phylogenetic tree of B cells
🅭🅯 This figure is distributed under the terms of the CC-BY licence
(`https://creativecommons.org/licenses/by/4.0/`), adapted from
Hoehn, Pybus and Kleinstein (2022) 348
12.3 Phylogenetic tree of single cells traced with a barcode
🅭🅯 This figure is distributed under the terms of the CC-BY licence
(`https://creativecommons.org/licenses/by/4.0/`), adapted from
Choi et al. (2022) 350

List of Tables

2.1 Three generations of sequencing technologies 31

4.1 Contingency table for GWAS example 71
4.2 Expected contingency table for GWAS example 72

5.1 Overview of substitution rate models and their parameters 111
5.2 Distance estimators for different nucleotide substitution models 122

6.1 Number of unrooted trees 144
6.2 Multiple sequence alignment (MSA) example 146
6.3 Hamming distance matrix for the example MSA 146
6.4 JC69 distance matrix for the example MSA 147
6.5 UPGMA example: intermediate distance matrix after one iteration 149
6.6 UPGMA example: intermediate distance matrix after two iterations 150
6.7 UPGMA example: tree distance matrix for the reconstructed tree 151

7.1 Errors in statistical tests 190
7.2 AIC values of different substitution models for a plant dataset example . . . 192

8.1 Assessing discrete character associations: contingency table 203
8.2 Assessing discrete character associations: contingency table for changes per branch 205
8.3 Assessing life history trait associations in birds 219

List of Boxes

1 p-value . . . 24
2 Binomial coefficient . . . 25
3 Binomial distribution . . . 25
4 Sequencing glossary . . . 28
5 Polymerase chain reaction (PCR) . . . 29
6 Emulsion polymerase chain reaction . . . 35
7 Landau symbol and algorithmic runtimes . . . 55
8 Contingency table test . . . 75
9 Hypergeometric distribution . . . 76
10 Fisher's exact test . . . 77
11 χ^2 distribution . . . 78
12 Pearson's χ^2-test . . . 79
13 Normal distribution . . . 80
14 Γ distribution (gamma distribution) . . . 81
15 Geometric distribution . . . 86
16 Exponential function . . . 88
17 Exponential distribution . . . 89
18 Connection between exponential and geometric distribution . . . 90
19 Diagonalisable matrix . . . 94
20 Bernoulli process . . . 95
21 Poisson process . . . 96
22 Connection between Poisson and binomial distribution . . . 97
23 Relationships between different distributions . . . 98
24 Markov chain . . . 98
25 Maximum likelihood estimator (MLE) . . . 116
26 Confidence interval . . . 117
27 Statistical consistency . . . 152
28 NP-completeness and NP-hardness . . . 154
29 Maximum likelihood tree inference . . . 198
30 Brownian Motion . . . 209
31 Glossary of MCMC terms . . . 315

Index

R_0, **250**, 251–254, 271, 322
R_e, **271**, 283, 285, 322, 323
Γ distribution, **81**, 119, 120, 122, 192
Γ function, 78, **81**
χ^2 distribution, 72, **78**, 79, 117, 187, 188, 195
$\mathcal{O}$, 54, **55**, 56, 65, 85, 87, 91, 92, 143, 152–154, 161, 167, 169, 170, 175, 228, 230, 246, 279, 352
d_N/d_S ratio, **128**, 133
k-mer, **57**, 58, 64, 65
p-value, 23, **24**, 70–74, 77–79, 185, 202–204, 208, 217–219

alignment, 3, 5, 7, 8, 34, **45**, 46, 64, 69, 114, 115, 119, 124, 145
  see also MSA, *see also* pairwise alignment
allele, **11**, 12, 69–71, 74, 256, 327
  dominant allele, **12**
  recessive allele, **12**
  SNP-allele, **69**, 70, 74
amino acid, 15, 45, 47, 48, 58, 69, 83, 99, 111, 124–126, 129, 130, 145, 158, 164, 165
amplification, **28**, 31, 34–36, 40, 41
assembly, 5, 7, **28**, 41, 46, 64, 65

basic local alignment search tool, *see* BLAST
basic reproductive number, *see* R_0
Bayes' rule, **302**, 303, 304, 310
Bayesian inference, 8, 63, 135, 145, **164**, 233, 248, 250–253, 266, 270, 294, 296, 298, 301–304, 306, 307, 309–312, 319–323, 335, 337, 340–342, 346, 348, 350, 352, 353, 355
  credible interval, **307**
    central posterior density interval, **307**, 308
    credible region, **307**
    highest posterior density interval, *see* Bayesian inference, credible interval, HPD
    HPD, **307**, 308, 322, 323
  credible set, **307**
Bernoulli process, 94, **95**
Bernoulli trial, **25**, 86, 95, 98, 114
binomial coefficient, 24, **25**, 234, 261, 291, 292, 305
binomial distribution, 24, **25**, 94, 95, 97, 115, 116, 288
birth-death model, **224**, 248, 249, 251, 256, 262, 267–269, 272, 273, 275, 296, 309, 322, 355
  birth-death skyline plot, **248**, 252, 271, 275, 298, 322
  constant rate model, **224**, 225–227, 232–236, 240, 242–248, 250, 251
  fossilised birth-death model, **324**, 325
  multi-type birth-death model, **250**, 277, 278, 280–283, 285, 286, 294, 297, 298, 354
BLAST, 5, 46, 48, 56, **57**, 58–60, 64
Bonferroni correction, 73
bootstrap, 182, **195**, 301
  bootstrap support, **198**, 199

non-parametric bootstrap, 194, **195**, 196–198, 251
parametric bootstrap, 194, **195**, 196, 253, 255
Brownian motion, 23, 208, **209**, 210–213, 216, 217, 219
brute-force, **158**, 165, 167, 168, 170

cell state transition diagram, **349**
central limit theorem, 80
character, **201**, 323
association, **201**, 205, 219
continuous character, **201**, 205, 208, 209, 213, 216–219
correlation, **201**, 202, 203, 205, 206, 208, 209, 213, 217, 220
discrete character, **201**, 202, 203, 205, 219
chi-squared distribution, *see* χ^2 distribution
cladistic approaches, 135, **145**
parsimony, **145**, 155, 162–164, 170, 173, 175, 178, 205, 347, 349, 350
Fitch algorithm, 156, **158**, 159, 161, 162, 167
maximum parsimony tree, **155**, 156, 157, 161, 170, 347
parsimony score, **155**, 156–162
coalescent model, 223, 224, **256**, 259–264, 267–270, 275, 296, 309, 333
coalescent limit, **259**, 263, 267, 288, 290, 332
coalescent skyline plot, **265**, 266, 270, 274, 298, 356
coalescent with gene conversion, **338**
coalescent with recombination, **334**, 335–338, 340
Kingman’s coalescent, **259**, 334
multiple sequentially Markovian coalescent, **338**
pairwise sequentially Markovian coalescent, **338**
structured coalescent model, **284**, 286, 287, 289–298, 354
structured Wright-Fisher process, **284**, 286–288, 296
Wright-Fisher process, **256**, 257–260, 263, 267, 269, 272, 273, 284, 333
Wright-Fisher process with recombination, **334**
codon, 15, 48, 83, 99, 111, 124, **126**, 127–133, 145, 158, 164, 165
start codon, **126**
stop codon, **126**
codon degeneracy, 15, **48**
compartmental model, **224**, 225
congruence class, **249**
contingency table, 71, 72, **75**, 77, 79, 185, 202, 203, 205
contrast method, **208**, 211, 216, 217
contrasts, **209**, 212
normalised contrasts, **209**, 213, 215–219
convergent evolution, 63, **144**, 163

decision problem, **154**
deoxyribonucleic acid, *see* DNA
detailed balance condition, *see* time-reversible process
difference equation, **92**, 227, 228
differential equation, **92**
distance-based methods, *see* phenetic approaches
DNA, 1, 5, 13, 14, 16–19, **27**, 28–31, 45, 47, 61, 69, 83, 84, 232, 323, 327, 328
DNA polymerase enzyme, **29**, 32, 35
DNA uptake, 327
dynamic programming, 48, **51**, 54, 57, 158, 167

Ebola virus, 17, 250, 251, 298, 322, 323, 345, 355
effective population size, **269**, 275, 292, 294, 295, 298, 309, 337, 338

effective reproductive number, *see* R_e
equilibrium frequencies, *see* stationary distribution
expectation maximisation, 338
exponential distribution, 87, **89**, 90, 94, 96, 225, 259, 260, 273, 292, 293, 305, 337
exponential function, 87, **88**, 89, 120, 260, 261

Felsenstein's pruning algorithm, **167**, 168–170, 178, 296
Fisher's exact test, 75, 76, **77**, 202, 203

Gamma distribution, *see* Γ distribution
Gaussian distribution, *see* normal distribution
gene, **11**, 12, 13, 15, 19, 27, 28, 56, 69, 73, 145, 182, 256, 269, 270
gene conversion, **330**, 331, 338
gene duplication, 45
gene flow, 13, 18, 338
genetic sequence, 5, 14, **27**, 45, 46, 56, 69, 84, 135, 137, 181, 295, 323, 324
  consensus sequence, **45**, 65, 181
  see also DNA, *see also* RNA
genome-wide association study, *see* GWAS
genotype, 7, **12**, 13–19, 69, 74, 75, 297
geometric distribution, **86**, 87, 89, 90, 94, 95, 231, 257, 259
GWAS, 7, 9, 57, **69**, 70–75, 185, 219, 220, 327, 347

Hamming distance, **112**, 113, 146–148
HCV, 269–271, 345
hepatitis C virus, *see* HCV
heuristic, **5**, 46, 48, 56, 57, 61, 62, 64, 175, 177
HGT, 9, 13, 327, 330, 331, 341, 346, 347
hill climbing, **175**, 176
homologous recombination, **328**, 330, 331
homologue, 5, **45**, 57, 63, 145, 333, 338
  orthologue, **45**, 145
  paralogue, **45**
horizontal gene transfer, *see* HGT
human immunodeficiency virus, *see* retrovirus, HIV
hypergeometric distribution, 71, 72, **76**, 77, 79, 185, 202–204, 219, 220

indel, **17**, 18, 48, 49, 63, 83
influenza, 17, 125, 283–286, 295, 297, 298, 320, 321, 339, 340, 345, 355

Landau symbol, *see* $\mathcal{O}$
least-squares dating, *see* LSD
likelihood, 49, **116**, 164, 186, 187, 191, 236, 243, 302, 304–306
lineages through time plot, *see* LTT plot
linear regression, 173, 174, **206**, 208, 209, 213, 217, 218, 243, 244, 247
  coefficient of determination, **207**
log-likelihood ratio, **117**, 186–189
long branch attraction, **163**, 164
LSD, **173**, 175
LTT plot, **236**, 237–244, 247

Manhattan plot, **73**
marginal likelihood, **302**, 303, 310, 311, 314
Markov chain, 22, 84, 95, 96, **98**, 99, 108–110, 124, 126, 209, 235, 259, 260, 286, 296, 311, 313, 323
  aperiodic Markov chain, **96**
  irreducible Markov chain, **96**
  Markov property, **98**
  time-homogeneous Markov chain, 95, **98**, 235
master equation, **92**, 228
matrix
  diagonal matrix, 93, **94**, 108, 110
  diagonalisable matrix, 93, **94**, 110

  identity matrix, 91, **94**
  invertible matrix, **94**
  symmetric matrix, 58, **94**, 108–110
maximum likelihood, 8, 49, 114, 115, 119, 133, **164**, 170, 175, 178, 179, 181, 236, 243, 247, 249, 292, 303, 347, 352
  CI, 114, 115, **117**, 118, 133, 179, 186, 188, 193–196, 198, 251, 302, 307, 322
  confidence interval, *see* CI
  confidence region, **117**, 187, 193
  maximum likelihood estimator, *see* maximum likelihood, MLE
  MLE, 115, **116**, 118, 119, 133, 153, 164, 186, 187, 193–196, 244, 247, 251–253, 255, 266, 302, 304, 322
  tree inference, 135, 145, **169**, 170, 173, 177, 178, 191, 197–199, 332
mean substitution rate, 102, 119, 121, 192, 252
meiosis, **329**, 330
memoryless process, 86, 95, **98**, 209
model adequacy, **355**
model selection, **185**, 189, 191, 193, 301, 308
molecular clock, **172**, 173, 174, 355
  relaxed clock, **172**
  strict clock, 148, **172**, 173–175
Monte Carlo algorithm, **312**, 352
  Markov chain Monte Carlo, *see* Monte Carlo algorithm, MCMC
  MCMC, 311, **312**, 313–317, 319–322, 324, 337, 340, 352
    acceptance probability, 314, **315**, 317
    burn-in, **315**, 316–318, 320
    effective sample size, *see* Monte Carlo algorithm, MCMC, ESS
    ESS, **315**, 317, 319, 320
    proposal distribution, 313, **315**, 316–320
    slow mixing, **315**, 318, 319
  Metropolis-Hastings algorithm, *see* Monte Carlo algorithm, MCMC
most recent common ancestor, *see* MRCA
MRCA, 45, 113, 135, 136, 140, 173, 212, **257**, 292, 293, 324
  GMRCA, **335**
  grand MRCA, *see* MRCA, GMRCA
MSA, 46, **61**, 62, 63, 83, 95, 112, 145–147, 155, 156, 159, 163–167, 170, 172, 185, 188, 197, 198, 244, 296, 306, 309–311, 323, 328, 331, 332, 335, 336, 338
  consistency-based alignment, **62**
  evolution-based alignment, **63**
  iterative alignment, **62**, 63
  progressive alignment, **62**, 63
  structural alignment, **63**
multiple sequence alignment, *see* MSA
mutation, 7, **17**, 18, 45, 61, 83, 84, 123, 124, 145, 274, 276, 277, 347, 349, 355

Newick format, **140**
normal distribution, 78, **80**, 117, 207, 209, 217
NP-complete, 61, **154**
  travelling salesman problem, **154**
NP-hard, 153, **154**, 170, 175
nucleotide
  purine, **13**, 14, 102, 122
  pyrimidine, **13**, 14, 102, 122
null hypothesis, **24**, 70, 71, 73, 74, 77, 79, 185, 188, 202–204, 208, 217
  null model, **185**, 186, 187, 190, 202, 203, 208, 212, 213, 220

odds ratio, **70**, 71, 72
optimisation problem, 153, **154**
outgroup, **172**

p-distance, **112**, 113
pairwise alignment, **46**, 47, 56, 61, 62, 158
dot-matrix method, 46, **48**
exhaustive method, 48, **50**
global alignment, **47**, 48
Needleman-Wunsch algorithm, 48, **51**, 52–56, 62
local alignment, **47**, 57
Smith-Waterman algorithm, 48, **51**, 54–56, 60, 61, 64
PCR, 28, **29**, 30–33, 41
emulsion PCR, 34, **35**, 36, 37
phenetic approaches, 135, **145**, 146, 178
algorithmic methods, **147**, 148
neighbour-joining, 62, 147, **148**, 351
UPGMA, 62, 147, **148**, 149–152, 154–157, 165, 172, 302, 320
distance matrix, 62, **146**, 147–154
optimality methods, **147**
least squares method, **153**, 154, 175
phenotype, 7, 8, 11, **12**, 13, 15, 16, 18, 19, 23, 69, 74, 75, 119, 121, 144, 274
phylodynamic likelihood, **244**, 245–251, 262, 265, 272, 282, 294
phylodynamic model
birth-death model, *see* birth-death model
coalescent model, *see* coalescent model
phylodynamics, 2, 3, **8**, 9, 51, 196, 223, 225, 244, 251, 252, 263, 270, 273, 277, 283, 299, 309, 319, 322, 345, 347, 351, 352
phylogenetic likelihood, 107, **164**, 165–171, 173, 175, 186, 244, 248
phylogenetic network, 1, 9, **327**, 328, 341
ancestral recombination graph, *see* phylogenetic network, ARG
ARG, **334**, 337–340
binary network, **327**
coalescence node, **328**
internal node, **328**
leaf, *see* phylogenetic network, tip
node degree, **327**
reticulation node, **328**, 329, 341, 342
rooted network, **328**
tip, **328**
unrooted network, **327**
phylogenetic reconstruction, 1–3, **7**, 8, 9, 51, 57, 61, 83, 106, 107, 125, 135, 138, 140, 144–146, 153, 164, 172, 178, 181, 185, 188, 198, 201, 223, 233, 249, 253, 273, 309, 312, 319, 320, 328, 331, 332, 351, 352
cladistic approaches, *see* cladistic approaches
phenetic approaches, *see* phenetic approaches
probabilistic approaches, *see* probabilistic approaches
phylogenetic tree, 1, 2, 7, 8, 22, 113, 135–138, **139**, 140, 144–146, 170, 172, 173, 179, 180, 182, 194, 198, 201, 204–213, 218, 223–226, 231, 232, 234, 236, 241, 243–245, 247, 250, 252, 253, 255–257, 259, 262, 263, 273, 275–277, 280, 296, 299, 301, 311, 321–325, 327, 328, 331, 334, 346–350, 353
binary tree, **138**
caterpillar tree, **139**
cherry, **139**, 140, 148, 151, 152, 163, 212–216
clade, **139**, 179, 180, 197, 207, 217, 250, 324
gene tree, **341**, 342
internal node, **135**, 139, 140, 158, 161, 165–167, 169, 171–174, 197–199, 209, 211, 234, 246,

247, 258, 260, 273, 274, 282

local tree, **331**, 332

monophyletic group, *see* phylogenetic tree, clade

multi-type tree, **280**, 286

branch-typed tree, **280**, 283, 284, 286, 290, 291, 294

sample-typed tree, **280**, 281

node degree, **138**

pendant branch, **139**, 140, 141, 181

root, **139**

rooted tree, **139**, 140, 141, 143, 155, 157–159, 165, 166, 169–173, 175, 328

sampled ancestor, **232**, 233

species tree, 136, 138, 144, 233, **341**, 347

time tree, **172**, 224

tip, 112, 113, **135**, 136, 139, 141–143, 152, 155, 158, 159, 167, 170, 171, 181, 211–213, 215, 216, 233, 244, 246, 251, 263, 276, 277, 281–283, 292, 347, 349, 353

topology, **139**, 173, 175, 191–193, 196, 198, 234, 235, 247, 250, 262, 275, 319, 321, 352

tree move, *see* tree move

ultrametric tree, **139**, 148–152, 220

unrooted tree, **139**, 140–144, 148, 155–157, 161, 162, 165–167, 170, 172, 175

phylogenetics, *see* phylogenetic reconstruction

phylogeny, **1**, 3, 4, 7, 8, 23, 45, 46, 57, 107, 111

see also phylogenetic network, *see also* phylogenetic tree

phylogeography, **296**, 297, 354, 356

Poisson distribution, 94, **96**, 97, 288

Poisson process, 94, **96**, 98, 175, 225, 293

polymerase chain reaction, *see* PCR

population dynamics, **1**, 2–4, 135, 223, 224, 232, 235, 256, 263–266, 272, 274, 299, 309, 337, 338, 353

posterior, 164, 286, **302**, 304–308, 310–315, 319–324, 335, 337, 338, 340, 352, 355

posterior predictive distribution, **355**

primer, **28**, 29, 30, 32, 36–38, 40

prior, **302**, 303–311, 336–338, 353–355

probabilistic approaches, 135, 145, **164**, 303

Bayesian inference, *see* Bayesian inference

maximum likelihood, *see* maximum likelihood, tree inference

probability density, **20**, 21, 80, 81, 87, 89, 90, 243, 245–249, 259–263, 280–283, 286, 291, 292, 306, 311

profile likelihood, **194**

proof by induction, **141**, 142, 151, 158, 230

pull-of-the-present, 238, **241**

pulley principle, **171**

push-of-the-past, 238, **239**

query sequence, **57**, 58–60, 64

random variable, **20**, 21, 22, 24, 25, 76–78, 80, 81, 86, 89, 90, 96, 116, 119, 185, 202, 210–212, 227, 236, 267, 303, 307

covariance, **21**, 211, 212, 215

mean, **20**, 21, 76, 78, 80, 81, 86, 89, 96, 119, 187, 207, 217, 236, 257, 288, 293, 298, 307, 312

standard deviation, **21**, 80, 213, 267

state space, **19**, 20, 21, 95, 98, 107, 108, 124, 209, 227, 310, 311, 313, 315, 317, 341, 352

variance, **20**, 21, 80, 96, 114, 175, 187, 207, 209, 211, 213,

215–217, 219, 243, 272, 288, 293, 298, 307, 312
rate heterogeneity, **119**, 120–122, 192
read, 5, 27, **28**, 30, 34, 36, 39, 41, 64, 65
recombination, 4, 7, 9, **17**, 18, 30, 46, 70, 74, 83, 220, 327, 334–336, 339, 340, 347
repeat, **17**, 18, 35, 37, 45, 48, 49, 65
retrovirus, 17
HIV, 8, 17, 121, 123, 125, 135, 137, 172, 179–183, 198, 199, 201, 273, 345, 347
SIV, 179–181, 198, 199
ribonucleic acid, *see* RNA
RNA, 1, 14, 16, 18, **27**, 47, 61, 83, 126, 269, 345, 346, 353
runtime, **55**, 60, 320, 321
exponential runtime, 51, 54, **55**, 61, 143, 153, 167, 170, 352
linear runtime, **55**, 65, 161, 170, 175
polynomial runtime, 51, 54, **55**, 56, 152, 154, 175

SARS-CoV-2, 137, 138, 251, 253, 254, 297, 322, 324, 345, 352
segregating site, **112**, 114, 119, 122
sequence library, 5, 46, **57**, 57, 60
sequencing, 5, **27**, 28
first generation sequencing, *see* sequencing, Sanger sequencing
high-throughput methods, *see* sequencing, NGS
next-generation sequencing, *see* sequencing, NGS
NGS, 31, **34**, 35–37, 39, 41
Illumina sequencing, 34, **37**, 40
Ion Torrent sequencing, 34, **37**, 39
pyrosequencing, 34, **35**, 36
SOLiD sequencing, 34, **36**, 38
Sanger sequencing, **29**, 30, 31, 33, 34, 65
second generation sequencing, *see* sequencing, NGS
sequencing error, **28**, 29, 34, 41
single-molecule sequencing, 31, **40**
nanopore sequencing, **42**
PacBio sequencing, **41**, 42
third generation sequencing, *see* sequencing, single-molecule sequencing
severe acute respiratory syndrome coronavirus 2, *see* SARS-CoV-2
single nucleotide polymorphism, *see* SNP
SNP, **69**, 70–74, 185, 220, 331
linked SNPs, **329**
unlinked SNPs, 220, **329**, 331, 347
stationary distribution, 95, 96, **98**, 99–101, 104, 125, 320
statistical consistency, **152**, 153, 170
statistical test, **24**, 75, 78, 79, 185, 186, 190, 202
accuracy, **190**
AIC, 186, **191**, 192, 193, 196, 198, 301
Akaike information criterion *see* statistical test, AIC
Bayes factor, **309**
likelihood ratio test *see* statistical test, LRT
LRT, 117, **186**, 187, 189, 190, 192–196, 198
nested models, **186**, 301
Pearson's χ^2-test, 72, 75, **79**
power, **190**, 191
rejection threshold, **24**, 73, 74, 79, 185, 187, 189, 190
significance, **24**, 73, 74, 187, 190, 193, 202–204
test statistic, **186**
type I error, **190**
type II error, **190**, 191, 217
Stirling's approximation, 143
stochastic process, 20, 22, 23, 49, 94, 95, **98**, 107, 185, 209, 210, 219, 227, 244, 260, 267, 273, 284, 290, 352
substitution, **84**

back substitution, **113**
convergent substitution, **113**
multiple substitution, **112**
non-synonymous substitution, **126**, 127–132
parallel substitution, **112**
synonymous substitution, **126**, 127–132
transition, **102**, 104, 105, 111, 122, 126, 127
transversion, **102**, 104, 105, 111, 122, 126, 127, 133
substitution model, **84**
amino acid model, 58, **124**, 125
codon model, **111**, 126, 127
F81, **104**, 110, 189
GTR, **105**, 106, 110, 111, 121, 164, 189
HKY, **105**, 106, 110, 111, 122, 146, 164, 189, 192, 252
HKY+Γ, **119**, 192
JC69, **99**, 100–104, 106, 110–115, 118–120, 122, 125, 129, 132, 133, 146, 147, 164, 185, 186, 188, 189, 192, 194, 304, 305, 307
JC69+Γ, **119**, 120, 122, 192
K80, **102**, 103–105, 110, 111, 122, 169, 186, 188, 189, 192
K80+Γ, **119**, 122, 192
TN93, 104, **105**, 110, 111, 122
UNREST, **105**, 111, 121
substitution rate matrix, 84, **85**, 91, 93, 95, 98–100, 102–106, 109, 110, 124, 125, 127, 164, 165, 167, 185, 192, 244, 309, 320, 335

taxon, *see* phylogenetic tree, tip
template, 17, **28**, 29–32, 34–37, 39–43
template library, **28**, 41
time-reversible process, 105, **107**, 108–111, 124, 165, 170
transcription, **14**, 16, 17
reverse transcription, 5, **16**, 27, 28
transition probability matrix, 84, **85**, 86, 87, 91–93, 96, 98–104, 113, 119, 121, 124, 145, 166, 170
translation, **14**, 16, 126
transmission tree, **137**, 138, 172, 181, 182, 223, 274, 276, 322, 347, 353
tree move, **176**
nearest-neighbour interchange, *see* tree move, NNI
NNI, **176**, 319
SPR, **176**, 177, 319
subtree pruning and regrafting, *see* tree move, SPR
TBR, **176**, 177, 319
tree bisection and reconnection, *see* tree move, TBR

viral reassortment, 327, 331, 337, **339**, 340
virion, **18**, 27, 84, 121, 138, 201, 273, 274, 347

zoonosis, **179**, 180, 339

Bibliography

Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer.

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

Aldous, D. J. (2001). Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Statistical Science* **16**, 23–34.

Alemany, A., M. Florescu, C. Baron, J. Peterson-Maduro and A. van Oudenaarden. (2018). Whole-organism clone tracing using single-cell sequencing. *Nature* **556**, 108–112.

Altman, N. and M. Krzywinski. (2015). Association, correlation and causation. *Nature Methods* **12**, 899–900.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology* **215**, 403–410.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. H. Zhang, Z. Zhang, W. Miller and D. J. Lipman. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389–3402.

Altshuler, D., J. N. Hirschhorn, M. Klannemark et al. (2000). The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nature Genetics* **26**, 76–80.

Alves, J. M., T. Prieto and D. Posada. (2017). Multiregional tumor trees are not phylogenies. *Trends in Cancer* **3**, 546–550.

Amarasinghe, S. L., S. Su, X. Dong, L. Zappia, M. E. Ritchie and Q. Gouil. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology* **21**, 30.

Anderson, R. M. and R. M. May. (1979). Population biology of infectious diseases: Part I. *Nature* **280**, 361–367.

Andrews, C. A. (2010). Natural selection, genetic drift, and gene flow do not act in isolation in natural populations. *Nature Education Knowledge* **3**, 5.

Archibald, J. D. and D. H. Deutschman. (2001). Quantitative analysis of the timing of the origin and diversification of extant placental orders. *Journal of Mammalian Evolution* **8**, 107–124.

Arndt, P. F. and T. Hwa. (2005). Identification and measurement of neighbor-dependent nucleotide substitution processes. *Bioinformatics* **21**, 2322–2328.

Arthur, R. R., N. F. Hassan, M. Y. Abdallah, M. S. El-Sharkawy, M. D. Saad, B. G. Hackbart and I. Z. Imam. (1997). Hepatitis C antibody prevalence in blood donors in different governorates in Egypt. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **91**, 271–274.

Atmore, L. M., L. Martínez-García, D. Makowiecki, C. André, L. Lõugas, J. H. Barrett and B. Star. (2022). Population dynamics of Baltic herring since the Viking Age revealed by ancient DNA and genomics. *Proceedings of the National Academy of Sciences* **119**, e2208703119.

Baltimore, D. (1970). RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature* **226**, 1209–1211.

Bankevich, A., S. Nurk, D. Antipov et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* **19**, 455–477.

Barido-Sottani, J., V. Bošková, L. du Plessis et al. (2018). Taming the BEAST — A community teaching material resource for BEAST 2. *Systematic Biology* **67**, 170–174.

Barido-Sottani, J., T. G. Vaughan and T. Stadler. (2020). A multitype birth–death model for Bayesian inference of lineage-specific birth and death rates. *Systematic Biology* **69**, 973–986.

Barley, A. J., A. Nieto-Montes de Oca, N. L. Manríquez-Morán and R. C. Thomson. (2022). The evolutionary network of whiptail lizards reveals predictable outcomes of hybridization. *Science* **377**, 773–777.

Barrat-Charlaix, P., T. G. Vaughan and R. A. Neher. (2022). TreeKnit: Inferring ancestral reassortment graphs of influenza viruses. *PLoS Computational Biology* **18**, e1010394.

Barré-Sinoussi, F., J. C. Chermann, F. Rey et al. (1983). Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science* **220**, 868–871.

Barton, N., A. Etheridge and A. Véber. (2010). A new model for evolution in a spatial continuum. *Electronic Journal of Probability* **15**, 162–216.

Beaulieu, J. M. and B. C. O'Meara. (2016). Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. *Systematic Biology* **65**, 583–601.

Beerenwinkel, N., R. F. Schwarz, M. Gerstung and F. Markowetz. (2014). Cancer evolution: mathematical models and computational inference. *Systematic Biology* **64**, e1–e25.

Bertoin, J. (1994). *Lévy Processes*. Cambridge University Press.

Bininda-Emonds, O. R. P., M. Cardillo, K. E. Jones et al. (2007). The delayed rise of present-day mammals. *Nature* **446**, 507–512.

Bloomquist, E. W. and M. A. Suchard. (2010). Unifying vertical and nonvertical evolution: a stochastic ARG-based framework. *Systematic Biology* **59**, 27–41.

Bollback, J. P. (2002). Bayesian model adequacy and choice in phylogenetics. *Molecular Biology and Evolution* **19**, 1171–1180.

Bošková, V., S. Bonhoeffer and T. Stadler. (2014). Inference of epidemiological dynamics based on simulated phylogenies using birth-death and coalescent models. *PLOS Computational Biology* **10**, e1003913.

Bošková, V., T. Stadler and C. Magnus. (2018). The influence of phylodynamic model specifications on parameter estimates of the Zika virus epidemic. *Virus Evolution* **4**, vex044.

Bouchard-Côté, A. and M. I. Jordan. (2013). Evolutionary inference via the Poisson Indel Process. *Proceedings of the National Academy of Sciences* **110** (4), 1160–1166.

Bouckaert, R., P. Lemey, M. Dunn, S. J. Greenhill, A. V. Alekseyenko, A. J. Drummond, R. D. Gray, M. A. Suchard and Q. D. Atkinson. (2012). Mapping the origins and expansion of the Indo-European language family. *Science* **337**, 957–960.

Bouckaert, R. R. and J. Heled. (2014). DensiTree 2: seeing trees through the forest. *bioRxiv*.

Bouckaert, R. R., J. Heled, D. Kühnert, T. Vaughan, C.-H. Wu, D. Xie, M. A. Suchard, A. Rambaut and A. J. Drummond. (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Computational Biology* **10**, e1003537.

Bouckaert, R. R., T. G. Vaughan, J. Barido-Sottani et al. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Computational Biology* **15**, e1006650.

Bouckaert, R. R. and M. Robbeets. (2017). Pseudo Dollo models for the evolution of binary characters along a tree. *bioRxiv*.

Boussau, B. and M. Gouy. (2006). Efficient likelihood computations with nonreversible models of evolution. *Systematic Biology* **55**, 756–768.

Box, G. E. and G. C. Tiao. (1992). *Bayesian Inference in Statistical Analysis*. John Wiley & Sons, Ltd.

Břinda, K., A. Callendrello, K. C. Ma et al. (2020). Rapid inference of antibiotic resistance and susceptibility by genomic neighbour typing. *Nature Microbiology* **5**, 455–464.

Burnham, K. P. and D. R. Anderson. (2002). *Model Selection and Multimodel Inference*, 2nd edition. Springer.

Bush, W. S. and J. H. Moore. (2012). Chapter 11: Genome-wide association studies. *PLoS Computational Biology* **8**, e1002822.

Cannings, C. (1974). The latent roots of certain Markov chains arising in genetics: A new approach, I. Haploid models. *Advances in Applied Probability* **6**, 260–290.

Cariaso, M. and G. Lennon. (2012). SNPedia: a wiki supporting personal genome annotation, interpretation and analysis. *Nucleic Acids Research* **40** (issue D1), D1308–D1312.

Carstens, B. C., M. L. Smith, D. J. Duckett, E. M. Fonseca and M. T. C. Thomé. (2022). Assessing model adequacy leads to more robust phylogeographic inference. *Trends in Ecology & Evolution* **37**, 402–410.

Casadesús, J. and D. Low. (2006). Epigenetic gene regulation in the bacterial world. *Microbiology and Molecular Biology Reviews* **70**, 830–856.

Casasent, A. K., A. Schalck, R. Gao et al. (2018). Multiclonal invasion in breast tumors identified by topographic single cell sequencing. *Cell* **172**, 205–217.

Cavalli-Sforza, L. L. and A. W. F. Edwards. (1967). Phylogenetic analysis: models and estimation procedures. *Evolution* **21**, 550–570.

Charlesworth, B. (2009). Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics* **10**, 195–205.

Chatzou, M., C. Magis, J.-M. Chang, C. Kemena, G. Bussotti, I. Erb and C. Notredame. (2016). Multiple sequence alignment modeling: methods and applications. *Briefings in Bioinformatics* **17**, 1009–1023.

Check Hayden, E. (2014). Ebola declared a public-health emergency. *Nature*.

Chernoff, H. and E. L. Lehmann. (1954). The use of maximum likelihood estimates in χ^2 tests for goodness of fit. *The Annals of Mathematical Statistics* **25**, 579–586.

Choi, J., W. Chen, A. Minkina et al. (2022). A time-resolved, multi-symbol molecular recorder via sequential genome editing. *Nature* **608**, 98–107.

Choo, Q.-L., G. Kuo, A. J. Weiner, L. R. Overby, D. W. Bradley and M. Houghton. (1989). Isolation of a cDNA cLone derived from a blood-borne non-A, non-B viral hepatitis genome. *Science* **244**, 359–362.

Chow, K.-H. K., M. W. Budde, A. A. Granados et al. (2021). Imaging cell lineage with a synthetic digital recording system. *Science* **372**, eabb3099.

Cole, S. R., H. Chu and S. Greenland. (2014). Maximum likelihood, profile likelihood, and penalized likelihood: a primer. *American Journal of Epidemiology* **179**, 252–260.

Comeron, J. M. (1995). A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. *Journal of Molecular Evolution* **41**, 1152–1159.

Compeau, P. E. C., P. A. Pevzner and G. Tesler. (2011). How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology* **29**, 987–991.

Corder, E. H., A. M. Saunders, W. J. Strittmatter, D. E. Schmechel, P. C. Gaskell, G. W. Small, A. D. Roses, J. L. Haines and M. A. Pericak-Vance. (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261**, 921–923.

Croucher, N. J., A. J. Page, T. R. Connor, A. J. Delaney, J. A. Keane, S. D. Bentley, J. Parkhill and S. R. Harris. (2014). Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research* **43**, e15.

Currie, T. E., S. J. Greenhill, R. D. Gray, T. Hasegawa and R. Mace. (2010). Rise and fall of political complexity in island South-East Asia and the Pacific. *Nature* **467**, 801–804.

Danecek, P., J. K. Bonfield, J. Liddle et al. (2021). Twelve years of SAMtools and BCFtools. *GigaScience* **10**.

Dang, C. C., Q. S. Le, O. Gascuel and V. S. Le. (2010). FLU, an amino acid substitution model for influenza proteins. *BMC Evolutionary Biology* **10**, 1–11.

Darriba, D., G. L. Taboada, R. Doallo and D. Posada. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* **9**, 772.

Darwin, C. (1837). *Notebook B*. Cambridge University Library.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray.

Dawkins, R. (2009). *The Greatest Show on Earth: The Evidence for Evolution*. Free Press.

Day, W. H. (1987). Computational complexity of inferring phylogenies from dissimilarity matrices. *Bulletin of Mathematical Biology* **49**, 461–467.

Dayhoff, M. O., R. M. Schwartz and B. C. Orcutt (1978). *A model of evolutionary change in proteins*. In: *Atlas of Protein Sequence and Structure*. Vol. 5. National Biomedical Research Foundation.

de Bruijn, N. G. (1946). A combinatorial problem. *Proceedings of the Section of Sciences of the Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam* **49**, 758–764.

de Maio, N., C.-H. Wu, K. M. O'Reilly and D. Wilson. (2015). New routes to phylogeography: a Bayesian structured coalescent approximation. *PLoS Genetics* **11**, e1005421.

de Oliveira, T., O. G. Pybus, A. Rambaut et al. (2006). HIV-1 and HCV sequences from Libyan outbreak. *Nature* **444**, 836–837.

Degnan, J. H. and N. A. Rosenberg. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution* **24**, 332–340.

Dempster, A. P., N. M. Laird and D. B. Rubin. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **39**, 1–22.

Dhar, A., D. K. Ralph, V. N. Minin and F. A. Matsen I V. (2020). A Bayesian phylogenetic hidden Markov model for B cell receptor sequence analysis. *PLoS Computational Biology* **16**, e1008030.

Didelot, X., D. Lawson, A. Daarling and D. Falush. (2010). Inference of homologous recombination in bacteria using whole-genome sequences. *Genetics* **186**, 1435–1449.

Dinh, V., A. Bilge, C. Zhang and F. A. Matsen IV (2017). *Probabilistic path Hamiltonian Monte Carlo*. In: *Proceedings of the 34th International Conference on Machine Learning*. Vol. 70, 1009–1018.

Distefano, J. K. and D. M. Taverna. (2011). Technological issues and experimental design of gene association studies. *Methods in Molecular Biology* **700**, 3–16.

Do, C. B., M. S. P. Mahabhashyam, M. Brudno and S. Batzoglou. (2005). ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Research* **15**, 330–340.

Dobzhansky, T. (1937). *Genetics and the Origin of Species*. Columbia University Press.

Dobzhansky, T. (1973). Nothing in biology makes sense except in the light of evolution. *American Biology Teacher* **35**, 125–129.

Donnelly, P. and T. G. Kurtz. (1999). Genealogical processes for Fleming-Viot models with selection and recombination. *The Annals of Applied Probability* **9**, 1091–1148.

Doolittle, W. F. (1999). Phylogenetic classification and the universal tree. *Science* **284**, 2124–2128.

Dorey, F. (2010). In brief: the p value: what is it and what does it tell you? *Clinical Orthopaedics and Related Research* **468**, 2297–2298.

Drummond, A. J., S. Y. W. Ho, M. J. Phillips and A. Rambaut. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biology* **4**, e88.

Drummond, A. J. and A. Rambaut. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* **7**, 1.

Drummond, A. J., A. Rambaut, B. Shapiro and O. G. Pybus. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution* **22**, 1185–1192.

du Plessis, L. (2016). *Understanding the spread and adaptation of infectious diseases using genomic sequencing data*. PhD thesis. ETH Zürich.

Duchêne, D. A., S. Duchêne, E. C. Holmes and S. Y. W. Ho. (2015). Evaluating the Adequacy of Molecular Clock Models Using Posterior Predictive Simulations. *Molecular Biology and Evolution* **32**, 2986–2995.

Duchêne, S., R. Bouckaert, D. A. Duchêne, T. Stadler and A. J. Drummond. (2018). Phylodynamic model adequacy using posterior predictive simulations. *Systematic Biology* **68**, 358–364.

Edgar, R. C. and S. Batzoglou. (2006). Multiple sequence alignment. *Current Opinion in Structural Biology* **16**, 368–373.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792–1797.

Edwards, A. W. F. and L. L. Cavalli-Sforza. (1964). Reconstruction of evolutionary trees. *Phenetic and Phylogenetic Classification* **6**, 67–76.

Elloumi, M. (2017). *Algorithms for Next-Generation Sequencing Data: Techniques, Approaches, and Applications*. Springer.

Etienne, R. S., B. Haegeman, T. Stadler, T. Aze, P. N. Pearson, A. Purvis and A. B. Phillimore. (2011). Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. *Proceedings of the Royal Society B: Biological Sciences* **279**, 1300–1309.

Fang, W., C. M. Bell, A. Sapirstein, S. Asami, K. Leeper, D. J. Zack, H. Ji and R. Kalhor. (2022). Quantitative fate mapping: A general framework for analyzing progenitor state dynamics via retrospective lineage barcoding. *Cell* **185**, 4604–4620.

Faria, N. R., M. U. G. Kraemer, S. C. Hill et al. (2018). Genomic and epidemiological monitoring of yellow fever virus transmission potential. *Science* **361**, 894–899.

Farris, J. S. (1983). The logical basis of phylogenetic analysis. *Advances in Cladistics* **2**, 7–36.

Felsenstein, J. (1973). Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Zoology* **22**, 240–249.

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* **27**, 401–410.

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

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

Felsenstein, J. (1985). Phylogenies and the comparative method. *The American Naturalist* **125**, 1–15.

Felsenstein, J. (1988). Phylogenies and quantitative characters. *Annual Review of Ecology and Systematics* **19**, 445–471.

Felsenstein, J. (2003). *Inferring Phylogenies*. Sinauer Associates.

Felsenstein, J. (2012). A comparative method for both discrete and continuous characters using the threshold model. *The American Naturalist* **179**, 145–156.

Feng, J., W. S. Dewitt III, A. McKenna, N. Simon, A. D. Willis and F. A. Matsen IV. (2021). Estimation of cell lineage trees by maximum-likelihood phylogenetics. *The Annals of Applied Statistics* **15**, 343–362.

Ferragina, P. and G. Manzini (2000). *Opportunistic data structures with applications*. In: *Proceedings 41st Annual Symposium on Foundations of Computer Science*, 390–398.

Fisher, R. A. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society* **85**, 87–94.

Fisher, R. A. (1925). Applications of "Student's" distribution. *Metron* **5** (3), 90–104.

Fisher, R. A. (1956). *Mathematics of a lady tasting tea*. In: *The World of Mathematics*. Ed. by J. R. Newman. Vol. 3. Simon and Schuster.

Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. Clarendon Press.

Fitch, W. M. (1971). Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Biology* **20**, 406–416.

Fitch, W. M. and E. Margoliash. (1967). Construction of phylogenetic trees. *Science* **155**, 279–284.

FitzJohn, R. G., W. P. Maddison and S. P. Otto. (2009). Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. *Systematic Biology* **58**, 595–611.

Flicek, P. and E. Birney. (2009). Sense from sequence reads: methods for alignment and assembly. *Nature Methods* **6**, S6–S12.

Fonseca, E. M., D. J. Duckett, F. G. Almeida, M. L. Smith, M. T. C. Thomé and B. C. Carstens. (2022). Assessing model adequacy for Bayesian Skyline plots using posterior predictive simulation. *PLoS One* **17**, e0269438.

Ford, D., F. A. Matsen and T. Stadler. (2009). A method for investigating relative timing information on phylogenetic trees. *Systematic Biology* **58**, 167–183.

Forster, P. and C. Renfrew. (2006). *Phylogenetic Methods and the Prehistory of Languages*. McDonald Institute for Archeological Research.

Foulds, L. R. and R. L. Graham. (1982). The Steiner problem in phylogeny is NP-complete. *Advances in Applied Mathematics* **3**, 43–49.

Fraga, M. F., E. Ballestar, M. F. Paz et al. (2005). Epigenetic differences arise during the lifetime of monozygotic twins. *Proceedings of the National Academy of Sciences* **102**, 10604–10609.

Frank, C., M. K. Mohamed, G. T. Strickland et al. (2000). The role of parenteral antischistosomal therapy in the spread of hepatitis C virus in Egypt. *The Lancet* **355**, 887–891.

Gallo, R. C., P. S. Sarin, E. P. Gelmann et al. (1983). Isolation of human T-cell leukemia virus in acquired immune deficiency syndrome (AIDS). *Science* **220**, 865–867.

Gao, F., E. Bailes, D. L. Robertson et al. (1999). Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. *Nature* **397**, 436–441.

Garamszegi, L. Z. (2014). *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology*. Springer.

Gascuel, O. and A. McKenzie. (2004). Performance analysis of hierarchical clustering algorithms. *Journal of Classification* **21**, 3–18.

Gavryushkina, A., T. A. Heath, D. T. Ksepka, T. Stadler, D. Welch and A. J. Drummond. (2017). Bayesian total-evidence dating reveals the recent crown radiation of penguins. *Systematic Biology* **66**, 57–73.

Gavryushkina, A., D. Welch, T. Stadler and A. J. Drummond. (2014). Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. *PLoS Computational Biology* **10**, e1003919.

Gernhard, T. (2008). New analytic results for speciation times in neutral models. *Bulletin of Mathematical Biology* **70**, 1082–1097.

Gibbs, A. J. and G. A. McIntyre. (1970). The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences. *European Journal of Biochemistry* **16**, 1–11.

Gire, S. K., A. Goba, K. G. Andersen et al. (2014). Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* **345**, 1369–1372.

Girolami, M. and B. Calderhead. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**, 123–214.

Goldberg, E. E. and B. Igić. (2012). Tempo and mode in plant breeding system evolution. *Evolution* **66**, 3701–3709.

Goldberg, E. E., L. T. Lancaster and R. H. Ree. (2011). Phylogenetic inference of reciprocal effects between geographic range evolution and diversification. *Systematic Biology* **60**, 451–465.

Goldman, N. and Z. Yang. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* **11**, 725–736.

Grant, P. R., B. R. Grant, J. N. Smith, I. J. Abbott and L. K. Abbott. (1976). Darwin's finches: population variation and natural selection. *Proceedings of the National Academy of Sciences* **73**, 257–261.

Grenfell, B. T., O. G. Pybus, J. R. Gog, J. L. Wood, J. M. Daly, J. A. Mumford and E. C. Holmes. (2004). Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303**, 327–332.

Griffiths, R. C. and S. Tavaré. (1994). Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **344**, 403–410.

Grimm, D. G., D. Roqueiro, P. A. Salomé et al. (2017). easyGWAS: a cloud-based platform for comparing the results of genome-wide association studies. *The Plant Cell* **29**, 5–19.

Grimmett, G. and D. Stirzaker. (1992). *Probability and Random Processes*, 2nd edition. Clarendon Press.

Grollemund, R., S. Branford, K. Bostoen, A. Meade, C. Venditti and M. Pagel. (2015). Bantu expansion shows that habitat alters the route and pace of human dispersals. *Proceedings of the National Academy of Sciences* **112**, 13296–13301.

Guindon, S. and N. de Maio. (2021). Accounting for spatial sampling patterns in Bayesian phylogeography. *Proceedings of the National Academy of Sciences* **118**, e2105273118.

Guindon, S. and O. Gascuel. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* **52**, 696–704.

Guindon, S., H. Guo and D. Welch. (2016). Demographic inference under the coalescent in a spatial continuum. *Theoretical Population Biology* **111**, 43–50.

Hadfield, J., C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford and R. A. Neher. (2018). Nextstrain: real-time tracking of pathogen evolution **34**, 4121–4123.

Hahn, B. H., G. M. Shaw, K. M. de Cock and P. M. Sharp. (2000). AIDS as a zoonosis: scientific and public health implications. *Science* **287**, 607–614.

Haldane, J. B. (1932). *The Causes of Evolution*. Longmans, Green and Co.

Hansen, T. F. (1997). Stabilizing selection and the comparative analysis of adaptation. *Evolution* **51**, 1341–1351.

Harismendy, O., P. C. Ng, R. L. Strausberg et al. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology* **10**, R32.

Harmon, L. J. (2018). *Phylogenetic Comparative Methods: Learning from Trees*. CreateSpace Independent Publishing Platform.

Harvey, P. H., R. M. May and S. Nee. (1994). Phylogenies without fossils. *Evolution* **48**, 523–529.

Hasegawa, M., T. Yano and H. Kishino. (1984). A new molecular clock of mitochondrial-DNA and the veolution of Hominoids. *Proceedings of the Japan Academy, Series B* **60**, 95–98.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Heath, T. A. (2012). A hierarchical Bayesian model for calibrating estimates of species divergence times. *Systematic Biology* **61**, 793–809.

Heath, T. A., J. P. Huelsenbeck and T. Stadler. (2014). The fossilized birth–death process for coherent calibration of divergence-time estimates. *Proceedings of the National Academy of Sciences* **111**, E2957–E2966.

Hedge, J., S. J. Lycett and A. Rambaut. (2013). Real-time characterization of the molecular epidemiology of an influenza pandemic. *Biology Letters* **9**, 20130331.

Hein, J., M. H. Schierup and C. Wiuf. (2005). *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press.

Heled, J. and A. J. Drummond. (2008). Bayesian inference of population size history from multiple loci. *BMC Evolutionary Biology* **8**, 289.

Henikoff, S. and J. G. Henikoff. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences* **89**, 10915–10919.

Heydari, M., G. Miclotte, P. Demeester, Y. Van de Peer and J. Fostier. (2017). Evaluation of the impact of Illumina error correction tools on de novo genome assembly. *BMC Bioinformatics* **18**, 374.

Higgins, D. G. and P. M. Sharp. (1988). CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* **73**, 237–244.

Hinchliff, C. E., S. A. Smith, J. F. Allman et al. (2015). Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proceedings of the National Academy of Sciences* **112**, 12764–12769.

Hodcroft, E. B., M. Zuber, S. Nadeau et al. (2021). Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *Nature* **595**, 707–712.

Hoehn, K. B., A. Fowler, G. Lunter and O. G. Pybus. (2016). The diversity and molecular evolution of B-cell receptors during infection. *Molecular Biology and Evolution* **33**, 1147–1157.

Hoehn, K. B., O. G. Pybus and S. H. Kleinstein. (2022). Phylogenetic analysis of migration, differentiation, and class switching in B cells. *PLoS Computational Biology* **18**, e1009885.

Hoehn, K. B., J. A. Vander Heiden, J. Q. Zhou, G. Lunter, O. G. Pybus and S. H. Kleinstein. (2019). Repertoire-wide phylogenetic models of B cell molecular evolution reveal evolutionary signatures of aging and vaccination. *Proceedings of the National Academy of Sciences* **116**, 22664–22672.

Hoehn, K. B., G. Lunter and O. G. Pybus. (2017). A phylogenetic codon substitution model for antibody lineages. *Genetics* **206**, 417–427.

Hogeweg, P. and B. Hesper. (1984). The alignment of sets of sequences and the construction of phyletic trees: An integrated method. *Journal of Molecular Evolution* **20**, 175–186.

Höhna, S., W. A. Freyman, Z. Nolen, J. P. Huelsenbeck, M. R. May and B. R. Moore. (2019). A Bayesian approach for estimating branch-specific speciation and extinction rates. *bioRxiv*.

Höhna, S., M. J. Landis, T. A. Heath, B. Boussau, N. Lartillot, B. R. Moore, J. P. Huelsenbeck and F. Ronquist. (2016). RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic Biology* **65**, 726–736.

Hooper, E. J. (1999). *The River: A Journey to the Source of HIV and AIDS*. Little, Brown and Co.

Hoscheit, P. and O. G. Pybus. (2019). The multifurcating skyline plot. *Virus Evolution* **5**, vez031.

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology* **23**, 183–201.

Huelsenbeck, J. P. and F. Ronquist. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755.

Huson, D. H., R. Rupp and C. Scornavacca. (2010). *Phylogenetic Networks: Concepts, Algorithms and Applications*. Cambridge University Press.

Huxley, J. (1942). *Evolution, the Modern Synthesis*. George Alien & Unwin Ltd.

Idury, R. M. and M. S. Waterman. (1995). A new algorithm for DNA sequence assembly. *Journal of Computational Biology* **2**, 291–306.

Ina, Y. (1995). New methods for estimating the numbers of synonymous and nonsynonymous substitutions. *Journal of Molecular Evolution* **40**, 190–226.

Jaffe, A. M. (2006). The millennium grand challenge in mathematics. *Notices of the AMS* **53**.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press.

Jeffreys, H. (1983). *Theory of Probability*, 3rd edition. Clarendon Press.

Jones, D. T., W. R. Taylor and J. M. Thornton. (1992). The rapid generation of mutation data matrices from protein sequences. *Bioinformatics* **8**, 275–282.

Jukes, T. H. and C. R. Cantor (1969). *Evolution of protein molecules*. In: *Mammalian Protein Metabolism*. Ed. by H. Munro. Academic Press, 21–132.

Kang, S., N. Borgsmüller, M. Valecha, M. Markowska, J. Kuipers, N. Beerenwinkel, D. Posada and E. Szczurek. (2023). DelSIEVE: joint inference of single-nucleotide variants, somatic deletions, and cell phylogeny from single-cell DNA sequencing data. *bioRxiv*.

Kang, S., N. Borgsmüller, M. Valecha et al. (2022). SIEVE: joint inference of single-nucleotide variants and cell phylogeny from single-cell DNA sequencing data. *Genome Biology* **23**, 248.

Katoh, K., K. Misawa, K.-i. Kuma and T. Miyata. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30**, 3059–3066.

Keeling, M. J. and P. Rohani. (2008). *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press.

Kelly, F. P. (1979). *Reversibility and Stochastic Networks*. Wiley.

Kendall, D. G. (1948). On the generalized “birth-and-death” process. *The Annals of Mathematical Statistics* **19**, 1–15.

Kester, L. and A. van Oudenaarden. (2018). Single-cell transcriptomics meets lineage tracing. *Cell Stem Cell* **23**, 166–179.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* **16**, 111–120.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature* **217**, 624–626.

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

Klein, R. J., C. Zeiss, E. Y. Chew et al. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385–389.

Kouyos, R. D., V. von Wyl, S. Yerly et al. (2010). Molecular epidemiology reveals long-term changes in HIV type 1 subtype B transmission in Switzerland. *Journal of Infectious Diseases* **201**, 1488–1497.

Kozlov, A., J. M. Alves, A. Stamatakis and D. Posada. (2022). CellPhy: accurate and fast probabilistic inference of single-cell phylogenies from scDNA-seq data. *Genome Biology* **23**, 37.

Kraemer, M. U. G., V. Hill, C. Ruis et al. (2021). Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B.1.1.7 emergence. *Science* **373**, 889–895.

Kubo, T. and Y. Iwasa. (1995). Inferring the rates of branching and extinction from molecular phylogenies. *Evolution* **49**, 694–704.

Kuhn, T. S., A. Ø. Mooers and G. H. Thomas. (2011). A simple polytomy resolver for dated phylogenies. *Methods in Ecology and Evolution* **2**, 427–436.

Kühnert, D., M. Coscolla, D. Brites, D. Stucki, J. Metcalfe, L. Fenner, S. Gagneux and T. Stadler. (2018). Tuberculosis outbreak investigation using phylodynamic analysis. *Epidemics* **25**, 47–53.

Kühnert, D., R. Kouyos, G. Shirreff et al. (2018). Quantifying the fitness cost of HIV-1 drug resistance mutations through phylodynamics. *PLoS Pathogens* **14**, e1006895.

Kühnert, D., T. Stadler, T. G. Vaughan and A. J. Drummond. (2016). Phylodynamics with migration: A computational framework to quantify population structure from genomic data. *Molecular Biology and Evolution* **33**, 2102–2116.

Kuo, G., Q.-L. Choo, H. J. Alter et al. (1989). An assay for circulating antibodies to a major etiologic virus of human non-A, non-B hepatitis. *Science* **244**, 362–364.

Lamarck, J.-B. (1809). *Philosophie Zoologique ou Exposition des Considérations relatives à l’histoire naturelle des Animaux*. Dentu et L’Auteur.

Langmead, B. and S. L. Salzberg. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359.

Lartillot, N. and H. Philippe. (2006). Computing Bayes factors using thermodynamic integration. *Systematic Biology* **55**, 195–207.

Le, S. Q. and O. Gascuel. (2008). An improved general amino acid replacement matrix. *Molecular Biology and Evolution* **25**, 1307–1320.

Legried, B. and J. Terhorst. (2022). A class of identifiable phylogenetic birth–death models. *Proceedings of the National Academy of Sciences* **119**, e2119513119.

Legried, B. and J. Terhorst. (2023). Identifiability and inference of phylogenetic birth-death models. *Journal of Teoretical Biology* **568**, 111520.

Lemey, P., A. Rambaut, T. Bedford et al. (2014). Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. *PLoS Pathogens* **10**, e1003932.

Lemey, P., A. Rambaut, A. J. Drummond and M. A. Suchard. (2009). Bayesian phylogeography finds its roots. *PLoS Computational Biology* **5**, e1000520.

Lemey, P., A. Rambaut, J. J. Welch and M. A. Suchard. (2010). Phylogeography takes a relaxed random walk in continuous space and time. *Molecular Biology and Evolution* **27**, 1877–1885.

Leventhal, G. E., H. F. Günthard, S. Bonhoeffer and T. Stadler. (2014). Using an epidemiological model for phylogenetic inference reveals density dependence in HIV transmission. *Molecular Biology and Evolution* **31**, 6–17.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993.

Li, H. and R. Durbin. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760.

Li, H. and R. Durbin. (2011). Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496.

Li, H., B. Handsaker, A. Wysoker et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079.

Li, H., J. Ruan and R. Durbin. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research* **18**, 1851–1858.

Li, R., Y. Li, K. Kristiansen and J. Wang. (2008). SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**, 713–714.

Li, R., C. Yu, Y. Li, T.-W. Lam, S.-M. Yiu, K. Kristiansen and J. Wang. (2009). SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966–1967.

Li, W. H. (1993). Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *Journal of Molecular Evolution* **36**, 96–99.

Ling, S., Z. Hu, Z. Yang et al. (2015). Extremely high genetic diversity in a single tumor points to prevalence of non-Darwinian cell evolution. *Proceedings of the National Academy of Sciences* **112**, E6496–E6505.

Linz, B., F. Balloux, Y. Moodley et al. (2007). An African origin for the intimate association between humans and Helicobacter pylori. *Nature* **445**, 915–918.

Liu, K., S. Raghavan, S. Nelesen, C. R. Linder and T. Warnow. (2009). Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science* **324**, 1561–1564.

Lodish, H. F., A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore and J. Darnell. (2000). *Molecular Cell Biology*, 4th edition. W. H. Freeman and Co.

Lotka, A. J. (1910). Contribution to the theory of periodic reactions. *The Journal of Physical Chemistry* **14**, 271–274.

Louca, S., A. McLaughlin, A. MacPherson, J. B. Joy and M. W. Pennell. (2021). Fundamental identifiability limits in molecular epidemiology. *Molecular Biology and Evolution* **38**, 4010–4024.

Louca, S. and M. W. Pennell. (2020). Extant timetrees are consistent with a myriad of diversification histories. *Nature* **580**, 502–505.

Loveless, T. B., C. K. Carlson, V. J. Hu, C. A. D. Helmy, G. Liang, M. Ficht, A. Singhai and C. C. Liu. (2021). Molecular recording of sequential cellular events into DNA. *bioRxiv*.

Lowen, A. C. (2018). It's in the mix: Reassortment of segmented viral genomes. *PLoS Pathogens* **14**, e1007200.

Löytynoja, A. (2014). *Phylogeny-aware alignment with PRANK*. In: *Multiple Sequence Alignment Methods*. Ed. by D. J. Russell. Humana Press, 155–170.

Löytynoja, A. and N. Goldman. (2005). An algorithm for progressive multiple alignment of sequences with insertions. *Proceedings of the National Academy of Sciences* **102**, 10557–10562.

Ma, K. C., T. D. Mortimer, M. A. Duckett, A. L. Hicks, N. E. Wheeler, L. Sánchez-Busó and Y. H. Grad. (2020). Increased power from conditional bacterial genome-wide association identifies macrolide resistance mutations in Neisseria gonorrhoeae. *Nature Communications* **11**, 5374.

MacKay, D. J. C. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press.

Maddison, W. P., P. E. Midford and S. P. Otto. (2007). Estimating a binary character's effect on speciation and extinction. *Systematic Biology* **56**, 701–710.

Magee, A., M. Karcher, F. A. Matsen IV and V. M. Minin. (2023). How trustworthy is your tree? Bayesian phylogenetic effective sample size through the lns of Monte Carlo error. *Bayesian Analysis*, 1–29.

Maliet, O., F. Hartig and H. Morlon. (2019). A model with many small shifts for estimating species-specific diversification rates. *Nature Ecology & Evolution* **3**, 1086–1092.

Malleret, M. M., M. D. Freire, P. Lemes, F. T. Brum, A. Camargo and L. Verrastro. (2022). Phylogeography and species delimitation of the Neotropical frog complex (Hylidae: Scinax granulatus). *Zoologica Scripta* **51**, 330–347.

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

Martin, T. E. (1995). Avian life history evolution in relation to nest sites, nest predation, and food. *Ecological Monographs* **65**, 101–127.

Martinez, P., D. Mallo, T. G. Paulson, X. Li, C. A. Sanchez, B. J. Reid, T. A. Graham, M. K. Kuhner and C. C. Maley. (2018). Evolution of Barrett's esophagus through space and time at single-crypt and whole-biopsy levels. *Nature Communications* **9**, 794.

Marx, J. L. (1984). Strong new candidate for AIDS agent. *Science* **224**, 475–477.

Matzke, N. J. (2016). The evolution of antievolution policies after Kitzmiller versus Dover. *Science* **351**, 28–30.

May, M. R. and C. J. Rothfels. (2023). Diversification models conflate likelihood and prior, and cannot be compared using conventional model-comparison tools. *Systematic Biology* **72**, 713–722.

McDonald, S. M., M. I. Nelson, P. E. Turner and J. T. Patton. (2016). Reassortment in segmented RNA viruses: mechanisms and outcomes. *Nature Reviews Microbiology* **14**, 448–460.

McKenna, A., G. M. Findlay, J. A. Gagnon, M. S. Horwitz, A. F. Schier and J. Shendure. (2016). Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353**, aaf7907.

McKenna, A. and J. A. Gagnon. (2019). Recording development with single cell dynamic lineage tracing. *Development* **146**, dev169730.

Mendel, G. (1866). Versuche über Pflanzenhybriden. *Verhandlungen des naturforschenden Vereines in Brünn* **4**, 3–47.

Meredith, R. W., J. E. Janečka, J. Gatesy et al. (2011). Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. *Science* **334**, 521–524.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* **21**, 1087–1092.

Metzker, M. L., D. P. Mindell, X.-M. Liu, R. G. Ptak, R. A. Gibbs and D. M. Hillis. (2002). Molecular evidence of HIV-1 transmission in a criminal case. *Proceedings of the National Academy of Sciences* **99**, 14292–14297.

Michener, C. D. and R. R. Sokal. (1957). A quantitative approach to a problem in classification. *Evolution*, 130–162.

Miescher-Rüsch, F. (1871). *Über die chemische Zusammensetzung der Eiterzellen*. August Hirschwald.

Mirarab, S., N. Nguyen, S. Guo, L.-S. Wang, J. Kim and T. Warnow. (2015). PASTA: Ultra-Large Multiple Sequence Alignment for Nucleotide and Amino-Acid sequences. *Journal of Computational Biology* **5**, 377–386.

Mitov, V., K. Bartoszek and T. Stadler. (2019). Automatic generation of evolutionary hypotheses using mixed Gaussian phylogenetic models. *Proceedings of the National Academy of Sciences* **116**, 16921–16926.

Moler, C. and C. Van Loan. (1978). Nineteen dubious ways to compute the exponential of a matrix. *SIAM Review* **20**, 801–836.

Moler, C. and C. Van Loan. (2003). Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review* **45**, 3–49.

Moore, B. R., S. Höhna, M. R. May, B. Rannala and J. P. Huelsenbeck. (2016). Critically evaluating the theory and performance of Bayesian analysis of macroevolutionary mixtures. *Proceedings of the National Academy of Sciences* **113**, 9569–9574.

Moran, P. A. P. (1962). *The statistical processes of evolutionary theory*. Clarendon Press.

Morlon, H., T. L. Parsons and J. B. Plotkin. (2011). Reconciling molecular phylogenies with the fossil record. *Proceedings of the National Academy of Sciences* **108**, 16327–16332.

Morlon, H., S. Robin and F. Hartig. (2022). Studying speciation and extinction dynamics from phylogenies: addressing identifiability issues. *Trends in Ecology & Evolution* **37**, 497–506.

Müller, N. F., G. Dudas and T. Stadler. (2019). Inferring time-dependent migration and coalescence patterns from genetic sequence and predictor data in structured populations. *Virus Evolution* **5**.

Müller, N. F., D. A. Rasmussen and T. Stadler. (2017). The structured coalescent and its approximations. *Molecular Biology and Evolution* **34**, 2970–2981.

Müller, N. F., U. Stolz, G. Dudas, T. Stadler and T. G. Vaughan. (2020). Bayesian inference of reassortment networks reveals fitness benefits of reassortment in human influenza viruses. *Proceedings of the National Academy of Sciences* **117**, 17104–17111.

Murtagh, F. (1984). Complexities of hierarchic clustering algorithms: State of the art. *Computational Statistics Quarterly* **1**, 101–113.

Nadeau, S. A., T. G. Vaughan, C. Beckmann et al. (2023). Swiss public health measures associated with reduced SARS-CoV-2 transmission using genome data. *Science Translational Medicine* **15**, eabn7979.

Nasrallah, C. A., D. H. Mathews and J. P. Huelsenbeck. (2010). Quantifying the impact of dependent evolution among sites in phylogenetic inference. *Systematic Biology* **60**, 60–73.

Nee, S., E. C. Holmes, R. M. May and P. H. Harvey. (1994). Extinction rates can be estimated from molecular phylogenies. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **344**, 77–82.

Nee, S., R. M. May and P. H. Harvey. (1994). The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **344**, 305–311.

Needleman, S. B. and C. D. Wunsch. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* **48**, 443–453.

Neher, R. A. and O. Hallatschek. (2013). Genealogies of rapidly adapting populations. *Proceedings of the National Academy of Sciences* **110**, 437–442.

Nei, M. and T. Gojobori. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution* **3**, 418–426.

Nelder, J. A. and R. W. M. Wedderburn. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)* **135**, 370.

Newey, W. K. and D. McFadden (1994). *Large sample estimation and hypothesis testing*. In: vol. 4. Elsevier, 2111–2245.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical, Physical and Engineering Sciences* **236**, 333–380.

Nicholls, G. K. and R. D. Gray. (2008). Dated ancestral trees from binary trait data and their application to the diversification of languages. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 545–566.

Nickle, D. C., L. Heath, M. A. Jensen, P. B. Gilbert, J. I. Mullins and S. L. Kosakovsky Pond. (2007). HIV-specific probabilistic models of protein evolution. *PloS one* **2**, e503.

Nielsen, R. and Z. Yang. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**, 929–936.

NIH National Cancer Institute. (2023). *Definition of a SNP*. URL: https://www.cancer.gov/publications/dictionaries/genetics-dictionary/def/snp (visited on 11/09/2023).

Notohara, M. (1990). The coalescent and the genealogical process in geographically structured population. *Journal of Mathematical Biology* **29**, 59–75.

Notredame, C., D. G. Higgins and J. Heringa. (2000). T-Coffee: a novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology* **302**, 205–217.

Novák, Á., I. Miklós, R. Lyngsø and J. Hein. (2008). StatAlign: an extendable software package for joint Bayesian estimation of alignments and evolutionary trees. *Bioinformatics* **24**, 2403–2404.

Orjuela-Sánchez, P., N. D. Karunaweera, M. da Silva-Nunes et al. (2010). Research article Single-nucleotide polymorphism, linkage disequilibrium and geographic structure in the malaria parasite Plasmodium vivax: prospects for genome-wide association studies. *BMC Genetics* **11**, 65.

Ou, C.-Y., C. A. Ciesielski, G. Myers et al. (1992). Molecular epidemiology of HIV transmission in a dental practice. *Science* **256**, 1165–1171.

Pagel, M. (1994). Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society of London, Series B* **255**, 37–45.

Pamilo, P. and N. O. Bianchi. (1993). Evolution of the Zfx and Zfy genes - rates and interdependence between the genes. *Molecular Biology and Evolution* **10**, 271–281.

Pannell, J. R. (2003). Coalescence in a metapopulation with recurrent local extinction and recolonization. *Evolution* **57**, 949–961.

Pearson, K. (1900). X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **50**, 157–175.

Pearson, T. A. and T. A. Manolio. (2008). How to interpret a genome-wide association study. *JAMA* **299**, 1335–1344.

Pečerska, J., M. Gil and M. Anisimova. (2021). Joint alignment and tree inference. *bioRxiv*.

Pečerska, J., D. Kühnert, C. J. Meehan, M. Coscollá, B. C. de Jong, S. Gagneux and T. Stadler. (2021). Quantifying transmission fitness costs of multi-drug resistant tuberculosis. *Epidemics* **36**, 100471.

Pépin, J. (2021). *The cut hunter*. In: *The Origins of AIDS*. Cambridge University Press, 62–82.

Pérez-Losada, M., M. Arenas, J. C. Galán, F. Palero and F. González-Candelas. (2015). Recombination in viruses: mechanisms, methods of study, and evolutionary consequences. *Infection, Genetics and Evolution* **30**, 296–307.

Pertsemlidis, A. and J. W. Fondon. (2001). Having a BLAST with bioinformatics (and avoiding BLASTphemy). *Genome Biology* **2** (reviews 2002), 1–10.

Pitman, J. (1999). Coalescents with multiple collisions. *Annals of Probability*, 1870–1902.

Popinga, A., T. Vaughan, T. Stadler and A. J. Drummond. (2015). Inferring epidemiological dynamics with Bayesian coalescent inference: the merits of deterministic and stochastic models. *Genetics* **199**, 595–607.

Posada, D. (2008). jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution* **25**, 1253–1256.

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

Power, R. A., J. Parkhill and T. de Oliveira. (2017). Microbial genome-wide association studies: lessons from human GWAS. *Nature Reviews Genetics* **18**, 41–50.

Pybus, O. G., A. J. Drummond, T. Nakano, B. H. Robertson and A. Rambaut. (2003). The epidemiology and iatrogenic transmission of hepatitis C virus in Egypt: a Bayesian coalescent approach. *Molecular Biology and Evolution* **20**, 381–387.

Pybus, O. G., A. Rambaut and P. H. Harvey. (2000). An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics* **155**, 1429–1437.

Quinti, I., E. Renganathan, E. E. Ghazzawi, M. Divizia, G. Sawaf, S. Awad, A. Pana' and G. Rocchi. (1995). Seroprevalence of HIV and HCV infections in Alexandria, Egypt. *Zentralblatt für Bakteriologie* **283**, 239–244.

Rabier, C.-E., V. Berry, M. Stoltz, J. D. Santos, W. Wang, J.-C. Glaszmann, F. Pardi and C. Scornavacca. (2021). On the inference of complex phylogenetic networks by Markov Chain Monte-Carlo. *PLoS Computational Biology* **17**, e1008380.

Rabosky, D. L. (2014). Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. *PLoS One* **9**, e89543.

Rabosky, D. L. and E. E. Goldberg. (2015). Model inadequacy and mistaken inferences of trait-dependent speciation. *Systematic Biology* **64**, 340–355.

Ragan, M. A., J. O. McInerney and J. A. Lake. (2009). The network of life: genome beginnings and evolution. *Philosophical Transactions of the Royal Society B* **364**, 2169–2175.

Raj, B., D. E. Wagner, A. McKenna, S. Pandey, A. M. Klein, J. Shendure, J. A. Gagnon and A. F. Schier. (2018). Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nature Biotechnology* **36**, 442–450.

Rambaut, A., T. T. Lam, L. M. Carvalho and O. G. Pybus. (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evolution* **2**.

Rasmussen, D. A., O. Ratmann and K. Koelle. (2011). Inference for nonlinear epidemiological models using genealogies and time series. *PLoS Computational Bbiology* **7**, e1002136.

Rasmussen, D. A. and T. Stadler. (2019). Coupling adaptive molecular evolution to phylodynamics using fitness-dependent birth-death models. *eLife* **8**, e45562.

Rasmussen, M. D., M. J. Hubisz, I. Gronau and A. Siepel. (2014). Genome-wide inference of ancestral recombination graphs. *PLoS Genetics* **10**, e1004342.

Ray, S. C., R. R. Arthur, A. Carella, J. Bukh and D. L. Thomas. (2000). Genetic epidemiology of hepatitis C virus throughout Egypt. *The Journal of Infectious Diseases* **182**, 698–707.

Redelings, B. D. (2014). Erasing errors due to alignment ambiguity when estimating positive selection. *Molecular Biology and Evolution* **31**, 1979–1993.

Redelings, B. D. (2021). BAli-Phy version 3: model-based co-estimation of alignment and phylogeny. *Bioinformatics* **37**, 3032–3034.

Redelings, B. D. and M. A. Suchard. (2005). Joint Bayesian estimation of alignment and phylogeny. *Systematic Biology* **54**, 401–418.

Redelings, B. D. and M. A. Suchard. (2007). Incorporating indel information into phylogeny estimation for rapidly emerging pathogens. *BMC Evolutionary Biology* **7**, 40–19.

Ridley, M. (1983). *The Explanation of Organic Diversity: The Comparative Method and Adaptations for Mating*. Clarendon Press.

Roberts, G. O. and J. S. Rosenthal. (2016). Complexity bounds for Markov chain Monte Carlo algorithms via diffusion limits. *Journal of Applied Probability* **53**, 410–420.

Roch, S. (2006). A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **3**, 92–94.

Ronikier, M., N. Kuzmanović, D. Lakušić, I. Stevanoski, Z. Nikolov and N. E. Zimmermann. (2023). High-mountain phylogeography in the Balkan Peninsula: isolation pattern in a species of alpine siliceous grasslands and its possible background. *Alpine Botany* **133**, 101–115.

Rose, M. R. and T. H. Oakley. (2007). The new biology: beyond the Modern Synthesis. *Biology Direct* **2**, 1–17.

Ross, M. G., C. Russ, M. Costello, A. Hollinger, N. J. Lennon, R. Hegarty, C. Nusbaum and D. B. Jaffe. (2013). Characterizing and measuring bias in sequence data. *Genome Biology* **14**, R51.

Ross, S. M. (1996). *Stochastic Processes*, 2nd edition. Wiley.

Rudin, W. (1976). *Principles of Mathematical Analysis*, 3rd edition. MacGraw-Hill.

Sägesser, J. (2010). *Evolutionary analysis of viral dynamics*. Supervisors: Tanja Stadler and Roger D. Kouyos, project report available from Tanja Stadler upon request. ETH Zürich.

Sagitov, S. (1999). The general coalescent with asynchronous mergers of ancestral lines. *Journal of Applied Probability* **36**, 1116–1125.

Saitou, N. and M. Nei. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**, 406–425.

Salsburg, D. (2002). *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. Henry Holt and Co.

San, J. E., S. Baichoo, A. Kanzi, Y. Moosa, R. Lessells, V. Fonseca, J. Mogaka, R. Power and T. de Oliveira. (2020). Current affairs of microbial genome-wide association studies: approaches, bottlenecks and analytical pitfalls. *Frontiers in Microbiology* **10**, 3119.

Sanders, P., K. Mehlhorn, M. Dietzfelbinger and R. Dementiev (2019). *Hash tables and associative arrays*. In: *Sequential and Parallel Algorithms and Data Structures: The Basic Toolbox*. Vol. 55. Springer, 117–151.

Sanger, F., S. Nicklen and A. R. Coulson. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences* **74**, 5463–5467.

Schena, M., D. Shalon, R. W. Davis and P. O. Brown. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470.

Scherer, S. W. and P. M. Visscher. (2016). *Genome-Wide Association Studies: From Polymorphism to Personalized Medicine*. Cambridge University Press.

Schiffels, S. and R. Durbin. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics* **46**, 919–925.

Schwartz, R. and A. A. Schäffer. (2017). The evolution of tumour phylogenetics: principles and practice. *Nature Reviews Genetics* **18**, 213–229.

Sciré, J., J. Barido-Sottani, D. Kühnert, T. G. Vaughan and T. Stadler. (2022). Robust phylodynamic analysis of genetic sequencing data from structured populations. *Viruses* **14**, 1648.

Seidel, S. and T. Stadler. (2022). TiDeTree: a Bayesian phylogenetic framework to estimate single-cell trees and population dynamic parameters from genetic lineage tracing data. *Proceedings of the Royal Society B* **289**, 20221844.

Self, S. G. and K.-Y. Liang. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* **82**, 605–610.

Shapiro, J. and D. Noble. (2021). What prevents mainstream evolutionists teaching the whole truth about how genomes evolve? *Progress in Biophysics and Molecular Biology* **165**, 140–152.

Sharp, P. M., E. Bailes, R. R. Chaudhuri, C. M. Rodenburg, M. O. Santiago and B. H. Hahn. (2001). The origins of acquired immune deficiency syndrome viruses: where and when? *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **356**, 867–876.

Sharp, P. M., D. L. Robertson and B. H. Hahn. (1995). Cross-species transmission and recombination of 'AIDS' viruses. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **349**, 41–47.

Sievers, F., A. Wilm, D. Dineen et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* **7**, 539.

Simpson, G. G. (1944). *Tempo and Mode in Evolution*. Columbia University Press.

Simpson, J. T., K. Wong, S. D. Jackman, J. E. Schein, S. J. Jones and I. Birol. (2009). ABySS: a parallel assembler for short read sequence data. *Genome Research* **19**, 1117–1123.

Skilling, J. (2006). Nested sampling for general Bayesian computation. *Bayesian Analysis* **1**, 833–859.

Smith, T. F. and M. S. Waterman. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology* **147**, 195–197.

Sokal, R. R. and F. J. Rohlf. (2012). *Biometry*, 4th edition. W. H. Freeman and Co.

Sokal, R. R. and C. D. Michener. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin* **28**, 1409–1438.

Solís-Lemus, C. and C. Ané. (2016). Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genetics* **12**, e1005896.

Sollis, E., A. Mosaku, A. Abid et al. (2022). The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Research* **51** (issue D1), D977–D985.

Sottoriva, A., H. Kang, Z. Ma et al. (2015). A Big Bang model of human colorectal tumor growth. *Nature Genetics* **47**, 209–216.

Soucy, S. M., J. Huang and J. P. Gogarten. (2015). Horizontal gene transfer: building the web of life. *Nature Reviews Genetics* **16**, 472–482.

Spanjaard, B., B. Hu, N. Mitic, P. Olivares-Chauvet, S. Janjuha, N. Ninov and J. P. Junker. (2018). Simultaneous lineage tracing and cell-type identification using CRISPR–Cas9-induced genetic scars. *Nature Biotechnology* **36**, 469–473.

Staden, R. (1979). A strategy of DNA sequencing employing computer programs. *Nucleic Acids Research* **6**, 2601–2610.

Stadler, T. (2009). On incomplete sampling under birth-death models and connections to the sampling-based coalescent. *Journal of Theoretical Biology* **261**, 58–66.

Stadler, T. (2010). Sampling-through-time in birth–death trees. *Journal of Theoretical Biology* **267**, 396–404.

Stadler, T. (2011). Mammalian phylogeny reveals recent diversification rate shifts. *Proceedings of the National Academy of Sciences* **108**, 6187–6192.

Stadler, T. (2013). How can we improve accuracy of macroevolutionary rate estimates? *Systematic Biology* **62**, 321–329.

Stadler, T. and S. Bonhoeffer. (2013). Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **368**, 20120198.

Stadler, T., R. Kouyos, V. von Wyl et al. (2011). Estimating the basic reproductive number from viral sequence data. *Molecular Biology and Evolution* **29**, 347–357.

Stadler, T., D. Kühnert, S. Bonhoeffer and A. J. Drummond. (2013). Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proceedings of the National Academy of Sciences* **110**, 228–233.

Stadler, T., D. Kühnert, D. A. Rasmussen and L. du Plessis. (2014). Insights into the early epidemic spread of Ebola in Sierra Leone provided by viral sequence data. *PLoS Currents* **6**.

Stadler, T., O. G. Pybus and M. P. H. Stumpf. (2021). Phylodynamics for cell biologists. *Science* **371**, eaah6266.

Stadler, T. and M. Steel. (2019). Swapping birth anddeath: symmetries and transformations in phylodynamic models. *Systematic Biology* **68**.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690.

Stolz, U., T. Stadler, N. F. Müller and T. G. Vaughan. (2021). Joint inference of migration and reassortment patterns for viruses with segmented genomes. *Molecular Biology and Evolution* **39**, msab342.

Stolz, U., T. Stadler and T. G. Vaughan. (2024). Integrating transmission dynamics and pathogen evolution through a Bayesian approach. *bioRxiv*.

Strimmer, K. and O. G. Pybus. (2001). Exploring the demographic history of DNA sequences using the Generalized Skyline Plot. *Molecular Biology and Evolution* **18**, 2298–2305.

Stucki, D., M. Ballif, T. Bodmer et al. (2015). Tracking a tuberculosis outbreak over 21 years: strain-specific single-nucleotide polymorphism typing combined with targeted whole-genome sequencing. *The Journal of Infectious Diseases* **211**, 1306–1316.

Suchard, M. A. and B. D. Redelings. (2006). BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics* **22**, 2047–2048.

Sulston, J. E., E. Schierenberg, J. G. White and J. N. Thomson. (1983). The embryonic cell lineage of the nematode Caenorhabditis elegans. *Developmental Biology* **100**, 64–119.

Tam, V., N. Patel, M. Turcotte, Y. Bossé, G. Paré and D. Meyre. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics* **20**, 467–484.

Tamura, K. and M. Nei. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution* **10**, 512–526.

Tan, G., M. Gil, A. P. Löytynoja, N. Goldman and C. Dessimoz. (2015). Simple chained guide trees give poorer multiple sequence alignments than inferred trees in simulation and phylogenetic benchmarks. *Proceedings of the National Academy of Sciences* **112**, E99–100.

Tavaré, S. (1986). *Some probabilistic and statistical problems in the analysis of DNA sequences*. In: *Some Mathematical Questions in Biology—DNA Sequence Analysis*. Ed. by R. M. Miura. American Mathematical Society, 57–86.

The CRyPTIC Consortium. (2022). Genome-wide association studies of global Mycobacterium tuberculosis resistance to 13 antimicrobials in 10,228 genomes identify new resistance mechanisms. *PLoS Biology* **20**, e3001755.

The Durban Declaration. (2000). *Nature* **406**, 15–16.

The Global Consortium for H5N8 and Related Influenza Viruses. (2016). Role for migratory wild birds in the global spread of avian influenza H5N8. *Science* **354**, 213–217.

The International HapMap Consortium. (2003). The international HapMap project. *Nature* **426**, 789–796.

Thomas, C. M. and K. M. Nielsen. (2005). Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature Reviews Microbiology* **3**, 711–721.

Thompson, E. A. (1975). *Human Evolutionary Trees*. Cambridge University Press.

Thompson, J. D., D. G. Higgins and T. J. Gibson. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**, 4673–4680.

To, T.-H., M. Jung, S. Lycett and O. Gascuel. (2016). Fast dating using least-squares criteria and algorithms. *Systematic Biology* **65**, 82–97.

Truman, K., T. G. Vaughan, A. Gavryushkin and A. Gavryushkina. (2024). The fossilised birth-death model is identifiable. *bioRxiv*.

Tuffley, C. and M. Steel. (1997). Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bulletin of Mathematical Biology* **59**, 581–607.

Turakhia, Y., B. Thornlow, A. S. Hinrichs, N. de Maio, L. Gozashti, R. Lanfear, D. Haussler and R. Corbett-Detig. (2021). Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nature Genetics* **53**, 809–816.

Tzeng, Y. H., R. Pan and W.-H. Li. (2004). Comparison of three methods for estimating rates of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution* **21**, 2290–2298.

Uffelmann, E., Q. Q. Huang, N. S. Munung, J. de Vries, Y. Okada, A. R. Martin, H. C. Martin, T. Lappalainen and D. Posthuma. (2021). Genome-wide association studies. *Nature Reviews Methods Primers* **1**, 1–21.

Van den Oord, E. J. C. G. (2008). Controlling false discoveries in genetic studies. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: the official publication of the International Society of Psychiatric Genetics* **147B**, 637–644.

Varsani, A., P. Lefeuvre, P. Roumagnac and D. Martin. (2018). Notes on recombination and reassortment in multipartite/segmented viruses. *Current Opinion in Virology* **33**, 156–166.

Vaughan, T. G., G. E. Leventhal, D. A. Rasmussen, A. J. Drummond, D. Welch and T. Stadler. (2019). Estimating epidemic incidence and prevalence from genomic data. *Molecular Biology and Evolution* **36**, 1804–1816.

Vaughan, T. G., D. Welch, A. J. Drummond, P. J. Biggs, T. George and N. P. French. (2017). Inferring ancestral recombination graphs from bacterial genomic data. *Genetics* **205**, 857–870.

Vaughan, T. G., D. Kühnert, A. Popinga, D. Welch and A. J. Drummond. (2014). Efficient Bayesian inference under the structured coalescent. *Bioinformatics* **30**, 2272–2279.

Vaughan, T. G., S. A. Nadeau, J. Sciré and T. Stadler. (2020). *Phylodynamic analyses of outbreaks in China, Italy, Washington State (USA), and the Diamond Princess*. URL: https://virological.org/t/phylodynamic-analyses-of-outbreaks-in-china-italy-washington-state-usa-and-the-diamond-princess/439 (visited on 15/04/2024).

Vaughan, T. G., J. Sciré, S. A. Nadeau and T. Stadler. (2024). Estimates of early outbreak-specific SARS-CoV-2 epidemiological parameters from genomic data. *Proceedings of the National Academy of Sciences* **121**, e2308125121.

Volterra, V. (1928). Variations and fluctuations of the number of individuals in animal species living together. *ICES Journal of Marine Science* **3**, 3–51.

Volz, E. M. (2012). Complex population dynamics and the coalescent under neutrality. *Genetics* **190**, 187–201.

Volz, E. M. and S. D. W. Frost. (2014). Sampling through time and phylodynamic inference with coalescent and birth-death models. *Journal of the Royal Society, Interface* **11**, 20140945.

Volz, E. M., S. L. Kosakovsky Pond, M. J. Ward, A. J. Leigh Brown and S. D. W. Frost. (2009). Phylodynamics of infectious disease epidemics. *Genetics* **183**, 1421–1430.

Vos, M. and X. Didelot. (2009). A comparison of homologous recombination rates in bacteria and archaea. *The ISME journal* **3**, 199–208.

Wagner, D. E. and A. M. Klein. (2020). Lineage tracing meets single-cell omics: opportunities and challenges. *Nature Reviews Genetics* **21**, 410–427.

Wakely, J. (2016). *Coalescent Theory: An Introduction*. Macmillan Learning.

Wang, L. and T. Jiang. (1994). On the complexity of multiple sequence alignment. *Journal of Computational Biology* **1**, 337–348.

Watson, J. D. and F. H. C. Crick. (1953). Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature* **171**, 737–738.

Watts, J., O. Sheehan, Q. D. Atkinson, J. Bulbulia and R. D. Gray. (2016). Ritual human sacrifice promoted and sustained the evolution of stratified societies. *Nature* **532**, 228–231.

Weismann, A. (1893). *The Germ-Plasm; a Theory of Heredity*. Charles Scribner's Sons.

Wen, D., Y. Yu and L. Nakhleh. (2016). Bayesian inference of reticulate phylogenies under the multispecies network coalescent. *PLoS Genetics* **12**, e1006006.

Whelan, S. and N. Goldman. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution* **18**, 691–699.

Whidden, C. and F. A. Matsen. (2015). Quantifying MCMC exploration of phylogenetic tree space. *Systematic Biology* **64**, 472–491.

WHO. (2023). *Hepatitis C Fact Sheet*. URL: https://www.who.int/news-room/fact-sheets/detail/hepatitis-c (visited on 06/10/2023).

Wibmer, C. K., J. N. Bhiman, E. S. Gray, N. Tumba, S. S. A. Karim, C. Williamson, L. Morris and P. L. Moore. (2013). Viral escape from HIV-1 neutralizing antibodies drives increased plasma neutralization breadth through sequential recognition of multiple epitopes and immunotypes. *PLoS Pathogens* **9**, 1–16.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics* **9**, 60–62.

Williams, D. (2001). *Weighing the Odds: A Course in Probability and Statistics*. Cambridge University Press.

World Health Organization. (2023). *HIV*. URL: https://www.who.int/data/gho/data/themes/hiv-aids (visited on 27/09/2023).

Wright, S. (1931). Evolution in Mendelian populations. *Genetics* **16**, 97–159.

Wright, S. (1955). *Classification of the factors of evolution*. In: *Cold Spring Harbor Symposia on Quantitative Biology*. Vol. 20, 16–24.

Yang, Z. (1994). Estimating the pattern of nucleotide substitution. *Journal of Molecular Evolution* **39**, 105–111.

Yang, Z. (1996). Maximum-likelihood models for combined analyses of multiple sequence data. *Journal of Molecular Evolution* **42**, 587–596.

Yang, Z. (2014). *Molecular Evolution – A Statistical Approach*. Oxford University Press.

Yermanos, A., V. Greiff, N. J. Krautler, U. Menzel, A. Dounas, E. Miho, A. Oxenius, T. Stadler and S. T. Reddy. (2017). Comparison of methods for phylogenetic B-cell lineage inference using time-resolved antibody repertoire simulations (AbSim). *Bioinformatics* **33**, 3938–3946.

Yu, Y., R. M. Barnett and L. Nakhleh. (2013). Parsimonious inference of hbridization in the presence of incomplete lineage sorting **62**, 738–751.

Zafar, H., A. Tzen, N. Navin, K. Chen and L. Nakhleh. (2017). SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models. *Genome Biology* **18**, 1–20.

Zerbino, D. R. and E. Birney. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* **18**, 821–829.

Zhai, W., T. K.-H. Lim, T. Zhang et al. (2017). The spatial organization of intra-tumour heterogeneity and evolutionary trajectories of metastases in hepatocellular carcinoma. *Nature Communications* **8**, 4565.

Zhang, C., H. A. Ogilvie, A. J. Drummond and T. Stadler. (2018). Bayesian inference of species networks from multilocus sequence data. *Molecular Biology and Evolution* **35**, 504–517.

Zhang, C., T. Stadler, S. Klopfstein, T. A. Heath and F. Ronquist. (2016). Total-evidence dating under the fossilized birth–death process. *Systematic Biology* **65**, 228–249.

Zharkikh, A. (1994). Estimation of evolutionary distances between nucleotide sequences. *Journal of Molecular Evolution* **39**, 315–329.

Zhou, X. and M. Stephens. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* **44**, 821–824.

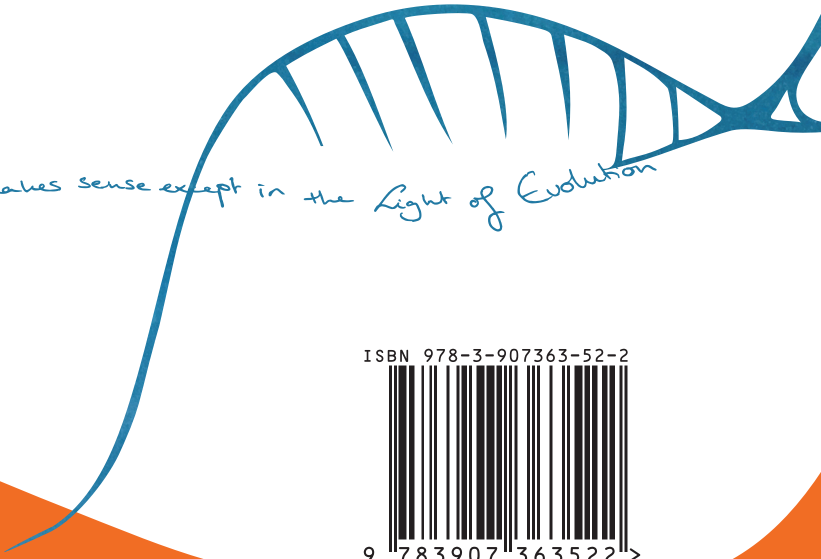
Zuckerkandl, E. and L. Pauling. (1962). Molecular disease, evolution, and genic heterogeneity. *Horizons in Biochemistry*, 189–225.

DECODING GENOMES

DECODING GENOMES demonstrates how to uncover information about past evolutionary and population dynamic processes based on genomic samples. The last decades have seen considerable theoretical and methodological advances in this area. These enable the assessment of critical scientific questions such as the impact of environmental changes on biodiversity and the evolution of pathogens during recent epidemics. The book gives the reader a detailed understanding of the whole process: from genome sampling to obtaining biological insights by applying sophisticated statistical and computational analyses. In particular, sequencing of genomic samples, the alignment of sequences, molecular evolution models, phylogenetics, and phylodynamics are core topics. Statistical and computational approaches discussed include dynamic programming, maximum likelihood, Bayesian statistics, and model selection, to name a few. The concepts introduced and applied throughout the book enable readers to answer questions across biological scales, including microevolution, macroevolution, immunology, development, cancer, and epidemiology, as well as in fields other than biology where evolutionary concepts are key, such as linguistics.

Target audience. The book is for students and researchers who aim to analyse genomic sequence data or develop statistical and computational approaches for such analyses. The content is tailored to readers from a wide variety of backgrounds, ranging from mathematics and statistics, computer science, or physics to biology, and, more generally, the life sciences.

The authors. All authors are or have been part of the Computational Evolution group at ETH Zürich, which is widely recognised as one of the leading teams in developing evolutionary and population dynamic models for analysing genomic data. Their various backgrounds — including mathematics, computer science, physics, and biology — help make this work accessible to a broad audience.



Nothing in Biology makes sense except in the Light of Evolution

$$\alpha(x'|x) = 1 \wedge \frac{\pi(x')}{\pi(x)} \cdot \frac{q(x|x')}{q(x'|x)}$$

ISBN 978-3-907363-52-2