Chapter 7
Recommendations for Further Reading

> *. . . a lot of people prefer Bayesian support values to other measures of support, such as the bootstrap . . . because they have a likeable tendency to give you higher numbers, making you feel happier about your tree.*
>
> – Lindell Bromham[1]

This chapter is a guide for the readers wanting to delve deeper into some of the topics of previous chapters.

7.1 Molecular Phylogenetics Books

Section 7.1.1 describes books requiring no more mathematics preparation than the present book. Section 7.1.2 briefly comments on books for readers with more knowledge of mathematics.

7.1.1 Phylogenetics Books with Less Mathematics

7.1.1.1 Phylogenetic Trees Made Easy: A How-To Manual [53]

Hall [53] shows step by step how to use phylogenetics software, especially MEGA 7. Nearly, all the MEGA 7 instructions apply with little modification to MEGA X,

Electronic Supplementary Material The online version contains supplementary material available at (https://doi.org/10.1007/978-3-031-11958-3_7).

[1] *An Introduction to Molecular Evolution and Phylogenetics* (Oxford University Press)[26, p. 431]. © Lindell Bromham 2016. Reproduced with permission of the Licensor through PLSclear.

D. R. Bickel, *Phylogenetic Trees and Molecular Evolution*, SpringerBriefs in Systems Biology, https://doi.org/10.1007/978-3-031-11958-3_7

the version described in Kumar et al. [74] and Stecher et al. [108] and mentioned above in Sects. 3.3, 5.4, and 5.6.

In addition to serving as a software tutorial for the topics introduced above in Chaps. 3–5, Hall [53] exposes the readers to specialized topics such as evolutionary networks and the detection of selection pressures. Like the present book (see the Preface), Hall [53] keeps the mathematics simple, including equations when needed for understanding.

7.1.1.2 An Introduction to Molecular Evolution and Phylogenetics [26]

Bromham [26] places molecular phylogenetics in the context of background topics such as mutation, replication, genomics, genetics, and mechanisms of evolution. Bromham [26] makes heavy use of graphical explanations, much as does Chap. 1, above. Due to its warnings about uncertainty in the output of phylogenetics software, Bromham [26] is cited above in Chaps. 3–4 to motivate their corrections of unquantified uncertainty.

To appeal to a wide audience of biologists, Bromham [26] avoids formulas as a matter of principle. For example, Bromham [26, p. 430] translated Bayes's theorem to English sentences much as Laplace had translated probability formulas into French sentences [78]. The principle is not followed slavishly: Bromham [26, p. 418] resorts to an equation in the discussion of evolution rates.

7.1.2 Phylogenetics Books with More Mathematics

Yang [133] gives details of many statistical methods of analyzing sequence data for molecular phylogenetics. The equations and notation are complex enough for describing the methods without recourse to the idealized simplifications seen in Chaps. 3 and 5 above. The author nonetheless made special efforts to make the book accessible to biologists [133, Preface]. Many uncertainties involved in statistical inferences about molecular evolution are thoroughly discussed in Yang [133, chapters 10–11]. The emphasis is on maximum likelihood estimation and Bayesian inference.

Drummond and Bouckaert [42] specifically focus on Bayesian inference about molecular evolution. The authors are the leading developers of BEAST 2 [24], which is currently the most popular software suite dedicated to Bayesian phylogenetics. The work Drummond and Bouckaert [42] is organized into three parts:

(1) The "Theory" part, like much of Yang [133], will appeal to the readers comfortable with calculus, linear algebra, and mathematical notation.
(2) The "Practice" part assists biologists with the use of BEAST 2 without requiring knowledge of the more technical parts.

(3) The "Programming" part, going under the hood of BEAST 2, will interest the readers with coding skills.

An earlier guide to BEAST is chapter 18 in Salemi et al. [104], a book written by the experts in specific areas of molecular phylogenetics, including the preliminary steps of finding and aligning sequences. This chapter describes MrBayes, another software tool for Bayesian inference.

The work Nei and Kumar [92] is written by the creators of the MEGA software mentioned in Sect. 7.1.1.1. Clearly explaining many statistical tests and bootstrap methods, it remains widely cited. Another classic text is Felsenstein [46].

Chapters 13 and 14 of Ewens and Grant [43] describe statistical methods of extracting information on molecular evolution from biological sequence data. Those chapters complement Nei and Kumar [92] in large part by providing a more concise treatment of the topics. Previous chapters of Ewens and Grant [43] give an overview of other statistical methods of analyzing DNA and protein sequences, with an emphasis on the statistics behind BLAST theory. Those methods are used to select and align sequences before the tree estimation methods can be applied. For practical guidance in that use of BLAST, see Hall [53], the book recommended in Sect. 7.1.1.1.

Xia [132] explains much of the mathematics involved in methods of phylogenetic trees reconstructed from sequence data, with an emphasis on distance-based methods and maximum likelihood estimation. The work Xia [132, chapter 2] is cited above in Sect. 3.4.1 on alignment as a source of uncertainty.

7.2 Bioinformatics and Genomics Books

The introductory book by Lesk [82] sets molecular phylogenetics in the context of other methods of computational biology. It is cited in Sect. 3.1.2.1 on distance-based estimation. Lesk [82] displays and discusses the three sequences behind Fig. 3.7.

Abu-Jamous et al. [1] describe many methods of cluster analysis in the context of bioinformatics problems. Distance-based tree estimation, while mathematically a form of hierarchical cluster analysis, has an evolutionary interpretation when homology is accepted as a working hypothesis (Sect. 1.2.1).

Using dice games, Bickel [12] explains statistical methods of analyzing data from genome-wide association studies and from measurements of gene expression and related proteomics and metabolomics data. The empirical Bayes tools (mentioned in Sect. 5.4) primarily apply to simultaneously testing multiple hypotheses. Multiple testing occurs not only in the types of data used in that book but also in an adjusted bootstrap proportion [102] (cf. Sect. 4.1) and also more generally, as seen in Sect. 2.5. Empirical Bayes methods are designed to guard against false positives like those leading to the replication crisis in many fields of science [see 8]. Motivated by that problem, chapter 7 of Bickel [12] explains how such methods scale down to

testing a single hypothesis. The book uses a result from the use of confidence theory to propagate the uncertainty in estimating prior distributions [17].

7.3 Imprecise Probability Books

You may recall that Sects. 4.2.3, 4.3, and 5.5.2 explain how to correct a confidence level or posterior probability by multiplying it by the estimated probability that the prior distributions and other model assumptions are adequate. That correction factor, the proportion of uncertainty quantified by the models, is $100\% - u$, where u is the proportion unquantified uncertainty [19].

Strict Bayesians would raise an objection: if all the probabilities are multiplied by $100\% - u$, then the total probability is $100\% - u$ instead of 100%. True, but "it's not a bug, it's a feature" [31], for such estimates honestly reflect the extent of unquantified uncertainty. (A less conservative method [15] is summarized above in Exercise 6b of Sect. 4.4.)

While the corrected probability function cannot be a probability distribution, it qualifies mathematically a *lower probability* function in the theory of imprecise probability. The value of such a function may be interpreted as the sufficiency of the evidence (Appendix A) or as a lower bound on standard probability. Under the latter interpretation, u serves as a "discounting coefficient" in the linear-vacuous [122, §2.9.2] or ε-contamination model of uncertainty described by Augustin et al. [6, §4.7]. Discounting coefficients that do not generate lower probabilities have also been considered [13, 14].

With that in mind, these books on imprecise probability theory are recommended to statisticians and scientists not averse to theorems:

- The work Augustin et al. [6], a unified collection of chapters by various experts, is mentioned out of chronological order since it provides an accessible entry to the topic.
- Walley [122] launched the field, providing mathematical and intuitive arguments for representing uncertainty with imprecise probability.
  - Walley [123] presents later developments by the same author.
- The work Troffaes and de Cooman [118] includes many of the theorems in Walley [122] as well as more recent results.
- 2021 saw the publication of two books with unique perspectives on imprecise probability:
  - Cuzzolin [37] not only presents two decades of research on a geometric approach but also reviews all major flavors of imprecise probability, citing over 2000 works. Discounting is discussed briefly [37, §4.3.6].
  - Weirich [126] puts emphasis on utility functions.

7.4 Power Law Books

Power laws form the core of the big-picture models of self-similar fluctuations seen in Appendices B-C. Such models have been used to describe fluctuating rates of molecular evolution (Sect. 2.8). The recommended starting point is Taleb [111], which offers a lively introduction to power laws.

Lowen and Teich [84] use fractal stochastic processes to define point processes in order to model count data such as the substitutions plotted in Fig. 1.1. A special case of such point processes is the class of intermittent point processes of Appendix B, below. Appendix C provides examples of other cases (Sect. C.2).

West et al. [129] introduce much of the mathematical modeling sketched in Sect. C.1.