

Contents

<i>Preface</i>	page xiii
<i>Glossary</i>	xvii
<i>Notation</i>	xviii

PART I BASIC TECHNIQUES 1

1	Brief Introduction to Phylogenetic Estimation	3
1.1	The Cavender–Farris–Neyman Model	4
1.2	An Analogy: Determining Whether a Coin is Biased Toward Heads or Tails	6
1.3	Estimating the Cavender–Farris–Neyman Tree	7
1.4	Some Comments about the CFN Model	16
1.5	Phylogeny Estimation Methods Used in Practice	16
1.6	Measuring Error Rates on Simulated Datasets	18
1.7	Getting Branch Support	20
1.8	Using Simulations to Understand Methods	20
1.9	Genome-Scale Evolution	23
1.10	Designing Methods for Improved Accuracy and Scalability	24
1.11	Summary	24
1.12	Review Questions	26
1.13	Homework Problems	27
2	Trees	29
2.1	Introduction	29
2.2	Rooted Trees	29
2.3	Unrooted Trees	35
2.4	Constructing the Strict Consensus Tree	41
2.5	Quantifying Error in Estimated Trees	41
2.6	The Number of Binary Trees on n Leaves	43
2.7	Rogue Taxa	43
2.8	Difficulties in Rooting Trees	44

2.9	Homeomorphic Subtrees	45
2.10	Some Special Trees	45
2.11	Further Reading	46
2.12	Review Questions	47
2.13	Homework Problems	47
3	Constructing Trees from True Subtrees	51
3.1	Introduction	51
3.2	Tree Compatibility	51
3.3	The Algorithm of Aho, Sagiv, Szymanski, and Ullman: Constructing Rooted Trees from Rooted Triples	52
3.4	Constructing Unrooted Binary Trees from Quartet Subtrees	53
3.5	Testing Compatibility of a Set of Trees	56
3.6	Further Reading	57
3.7	Review Questions	58
3.8	Homework Problems	58
4	Constructing Trees from Qualitative Characters	61
4.1	Introduction	61
4.2	Terminology	62
4.3	Tree Construction Based on Maximum Parsimony	63
4.4	Constructing Trees from Compatible Characters	69
4.5	Tree Construction Based on Maximum Compatibility	72
4.6	Treatment of Missing Data	75
4.7	Informative and Uninformative Characters	75
4.8	Further Reading	77
4.9	Review Questions	78
4.10	Homework Problems	78
5	Distance-based Tree Estimation Methods	83
5.1	Introduction	83
5.2	UPGMA	84
5.3	Additive Matrices	86
5.4	Estimating Four-Leaf Trees: The Four Point Method	87
5.5	Quartet-based Methods	89
5.6	Neighbor Joining	91
5.7	Distance-based Methods as Functions	92
5.8	Optimization Problems	94
5.9	Minimum Evolution	95
5.10	The Safety Radius	96
5.11	Comparing Methods	99
5.12	Further Reading	100
5.13	Review Questions	103
5.14	Homework Problems	104

6	Consensus and Agreement Trees	109
6.1	Introduction	109
6.2	Consensus Trees	109
6.3	Agreement Subtrees	116
6.4	Clustering Sets of Trees	117
6.5	Further Reading	117
6.6	Review Questions	118
6.7	Homework Problems	118
7	Supertrees	121
7.1	Introduction	121
7.2	Compatibility Supertrees	123
7.3	Asymmetric Median Supertrees	123
7.4	Robinson-Foulds Supertrees	124
7.5	Matrix Representation with Parsimony	126
7.6	Matrix Representation with Likelihood	128
7.7	Quartet-based Supertrees	128
7.8	The Strict Consensus Merger	132
7.9	SuperFine: A Meta-Method to Improve Supertree Methods	135
7.10	Further Reading	139
7.11	Review Questions	142
7.12	Homework Problems	142
	PART II MOLECULAR PHYLOGENETICS	143
8	Statistical Gene Tree Estimation Methods	145
8.1	Introduction to Statistical Estimation in Phylogenetics	145
8.2	Models of Site Evolution	146
8.3	Model Selection	151
8.4	Distance-based Estimation	152
8.5	Calculating the Probability of a Set of Sequences on a Model Tree	154
8.6	Maximum Likelihood	157
8.7	Bayesian Phylogenetics	159
8.8	Statistical Properties of Maximum Parsimony and Maximum Compatibility	161
8.9	The Impact of Taxon Sampling on Phylogenetic Estimation	164
8.10	Estimating Branch Support	165
8.11	Beyond Statistical Consistency: Sample Complexity	167
8.12	Absolute Fast Converging Methods	167
8.13	Heterotachy and the No Common Mechanism Model	170
8.14	Further Reading	172
8.15	Review Questions	173

8.16	Homework Problems	174
9	Multiple Sequence Alignment	178
9.1	Introduction	178
9.2	Evolutionary History and Sequence Alignment	180
9.3	Computing Differences Between Two Multiple Sequence Alignments	180
9.4	Edit Distances and How to Compute Them	184
9.5	Optimization Problems for Multiple Sequence Alignment	190
9.6	Sequence Profiles	194
9.7	Profile Hidden Markov Models	198
9.8	Reference-based Alignments	204
9.9	Template-based Methods	205
9.10	Seed Alignment Methods	206
9.11	Aligning Alignments	207
9.12	Progressive Alignment	209
9.13	Consistency	212
9.14	Weighted Homology Pair Methods	213
9.15	Divide-and-Conquer Methods	214
9.16	Co-estimation of Alignments and Trees	215
9.17	Ensembles of HMMs	220
9.18	Consensus Alignments	224
9.19	Discussion	226
9.20	Further Reading	227
9.21	Review Questions	231
9.22	Homework Problems	231
10	Phylogenomics: Constructing Species Phylogenies from Multi-Locus Data	234
10.1	Introduction	234
10.2	The Multi-Species Coalescent Model (MSC)	235
10.3	Using Standard Phylogeny Estimation Methods in the Presence of ILS	238
10.4	Probabilities of Gene Trees under the MSC	239
10.5	Coalescent-based Methods for Species Tree Estimation	241
10.6	Improving Scalability of Coalescent-based Methods	253
10.7	Species Tree Estimation under Duplication and Loss Models	254
10.8	Constructing Trees in the Presence of Horizontal Gene Transfer	259
10.9	Phylogenetic Networks	260
10.10	Further Reading	268
10.11	Review Questions	272
10.12	Homework Problems	272
11	Designing Methods for Large-Scale Phylogeny Estimation	274
11.1	Introduction	274
11.2	Standard Approaches	274

11.3	Introduction to Disk-Covering Methods (DCMs)	279
11.4	DCMs that Use Distance Matrices	282
11.5	Tree-based DCMs	285
11.6	Recursive Decompositions of Triangulated Graphs	288
11.7	Creating Multiple Trees	288
11.8	DACTAL: A General Purpose DCM	289
11.9	Triangulated Graphs	293
11.10	Further Reading	296
11.11	Review Questions	297
11.12	Homework Problems	298
Appendix A	Primer on Biological Data and Evolution	299
Appendix B	Algorithm Design and Analysis	304
Appendix C	Guidelines for Writing Papers About Computational Methods	327
Appendix D	Projects	331
	References	339
	Index	376