

Species Tree Inference

A Guide to Methods and Applications

EDITED BY
LAURA S. KUBATKO AND
L. LACEY KNOWLES

PRINCETON UNIVERSITY PRESS
Princeton and Oxford

Contents

	<i>Preface</i>	xvii
	<i>Acknowledgments</i>	xix
	<i>List of Contributors</i>	xxi
CHAPTER 1	Introduction to Species Tree Inference	1
	1.1 Introduction	1
	1.2 Background and Terminology	2
	1.2.1 Definitions and Terminology	2
	1.2.2 An Introduction to the Multispecies Coalescent	5
	1.2.3 Data Types and Technologies for Generating Phylogenomic Data	6
	1.3 Overview of Current Methods for Species Tree Inference	9
	1.3.1 Controversies in the Estimation of Species Trees	11
	1.4 A Look to the Future	12
	1.4.1 Current Limitations and Future Prospects	12
	1.4.2 Beyond the Species Tree	13
	1.5 Organization of This Book	14
PART I	Analytical and Methodological Developments	15
CHAPTER 2	Large-Scale Species Tree Estimation	19
	2.1 Introduction	19
	2.2 Species Tree Estimation Methods Addressing ILS	21
	2.2.1 Overview	21
	2.2.2 Summary Methods	21
	2.2.3 Coestimation Methods	24
	2.2.4 Site Based Methods	26
	2.2.5 Evaluation of Branch Support in Species Trees	28
	2.3 Species Tree Estimation under GDL	29
	2.4 Parallel Implementations for Species Tree Estimation	30
	2.4.1 ASTRALMP	30
	2.4.2 Multilocus Species Tree Estimation Using Maximum Likelihood	31

2.5	Divide-and-Conquer Species Tree Estimation	33
2.5.1	Divide-and-Conquer Using Supertree Methods	34
2.5.2	Divide-and-Conquer Using Disjoint Tree Merger Methods	34
2.6	Choice of Method	36
2.6.1	Statistical Consistency	36
2.6.2	Empirical Performance	37
2.7	Summary, Challenges, and Future Directions	39
2.8	Appendix: Big-O Analysis	41
CHAPTER 3	Species Tree Estimation Using ASTRAL: Practical Considerations	43
3.1	Introduction	43
3.2	ASTRAL Algorithm	46
3.2.1	Motivation and History	46
3.2.2	ASTRAL Algorithm	47
3.2.3	Summary of Known Theoretical Results Related to ASTRAL	50
3.3	Accuracy	51
3.4	Running Time	54
3.5	Input to ASTRAL: Practical Considerations	54
3.5.1	Gene Tree Estimation	55
3.5.2	Filtering of Data	57
3.6	ASTRAL Output	61
3.6.1	Species Tree Topology and Its Quartet Score	61
3.6.2	Branch Lengths in Coalescent Units	61
3.6.3	Branch Support Using Local Posterior Probability (localPP)	64
3.7	Follow-up Analyses and Visualization	65
3.7.1	Tests for Polytomies	65
3.7.2	Per Branch Quartet Support (Measure of Discordance)	65
3.8	Conclusion	66
CHAPTER 4	Species Tree Estimation Using Site Pattern Frequencies	68
4.1	Introduction	68
4.2	Estimation of the Species Tree Topology Using SVDQuartets	69
4.2.1	Theoretical Basis	69
4.2.2	Accounting of Incomplete Lineage Sorting in SVDQuartets	74
4.2.3	Species Tree Inference: Quartet Sampling and Assembly	75
4.2.4	Algorithmic Details	76
4.2.5	Uncertainty Quantification	78
4.2.6	Application to Species Relationships among Gibbons	78
4.2.7	Properties of SVDQuartets	79
4.2.8	Recommendations for Using SVDQuartets	82
4.3	Estimation of Speciation Times	82
4.3.1	Theoretical Basis	83
4.3.2	Algorithmic Details	86

4.3.3	Uncertainty Quantification	86
4.3.4	Application to Species Relationships among Gibbons	87
4.3.5	Recommendations for Using Composite Likelihood Estimators of the Speciation Times	87
4.4	Conclusion and Future Work	87
CHAPTER 5	Practical Aspects of Phylogenetic Network Analysis Using PhyloNet	89
5.1	Introduction	89
5.2	Reading and Interpretation of a Phylogenetic Network	91
5.2.1	Phylogenetic Network Parameters and Their Identifiability	92
5.3	Heuristic Searches, Point Estimates, and Posterior Distributions, or, Why Am I Getting Different Networks in Different Runs?	92
5.4	Illustration of the Various Inference Methods in PhyloNet	96
5.4.1	Inference under the MDC Criterion	96
5.4.2	Maximum Likelihood Inference	98
5.4.3	Maximum Pseudolikelihood Inference	102
5.4.4	Bayesian Inference	103
5.4.5	Running Time	105
5.5	Analysis of Larger Data Sets	106
5.6	Comparison and Summarization of Networks	111
5.6.1	Displayed Trees	111
5.6.2	Backbone Networks	111
5.6.3	Tree Decompositions	112
5.6.4	Tripartitions	112
5.6.5	Major Trees	112
5.7	Reticulate Evolutionary Processes in PhyloNet	112
5.7.1	Analysis of Polyploids	114
5.8	Conclusions	117
	Notes	119
CHAPTER 6	Network Thinking: Novel Inference Tools and Scalability Challenges	120
6.1	Introduction: The Impact of Gene Flow	120
6.2	Trees versus Networks	122
6.3	Species Networks	124
6.3.1	Explicit versus Implicit Networks	126
6.3.2	Extended Parenthetical Format	127
6.3.3	Displayed Trees and Subnetworks	128
6.3.4	Comparison of Networks	128
6.4	Fast Reconstruction of Species Networks	129
6.4.1	Maximum Pseudolikelihood Estimation	130
6.4.2	Rooting of Semidirected Networks	136
6.4.3	Goodness of Fit Tools	139
6.4.4	Bootstrap Analysis	140

6.5	Appendix: Installation and Use of the PhyloNetworks Julia Package	143
6.5.1	Main Functions in PhyloNetworks	143
PART II	Empirical Inference	145
CHAPTER 7	Phylogenomic Conflict in Plants	149
7.1	Introduction	149
7.2	Two Examples of Gene Tree Conflict within Angiosperms	152
7.3	The Consequences of Gene Tree Conflict in Phylogenomics	154
7.3.1	Inference of Species Trees	154
7.3.2	Gene Duplication and Genome Duplication	157
7.3.3	Divergence Time and Comparative Analyses	158
7.4	Resolution of the Tree of Plant Life	160
CHAPTER 8	Hybridization in <i>lochroma</i>	161
8.1	Introduction	161
8.2	Methods	163
8.2.1	Study System	163
8.2.2	Experimental Design	165
8.2.3	Target Capture and Assembly	166
8.2.4	Detection of Patterns of Hybridization from Gene Tree Distributions	167
8.2.5	Testing of Hybridization in Empirical Data Sets	168
8.3	Results	168
8.3.1	Addition of Hybrid Taxa Increases Discordance and Decreases Tree-Like Signal	168
8.3.2	Tests of Hybridization Support Different Relationships than Expected	170
8.4	Discussion	172
8.4.1	Effects of Hybridization on Patterns of Gene Tree Discordance	172
8.4.2	Challenges in Determining the Exact Hybrid Relationships	172
8.4.3	Hybridization in <i>lochrominae</i>	173
8.5	Conclusions	174
CHAPTER 9	Hybridization and Polyploidy in <i>Penstemon</i>	175
9.1	Introduction	175
9.2	Approach	176
9.2.1	Calculation of Quartet Concordance Factors	177
9.2.2	Bootstrapping and Gene Tree Uncertainty	178
9.2.3	Validation of QCF Estimation	178
9.2.4	Implementation	179
9.3	Materials and Methods	179
9.3.1	Study System	179
9.3.2	Sample Collection, DNA Extraction, and Amplicon Sequencing	180

9.3.3	Species Tree Inference	181
9.3.4	Candidate Hybridization Events from Rooted Triples	181
9.3.5	Species Network Inference	182
9.4	Results	182
9.4.1	Nuclear Amplicon Data	182
9.4.2	Species Tree Inference	182
9.4.3	Tests for Hybridization and Species Network Inference	186
9.5	Discussion	186
9.5.1	Taxonomy of Subsections <i>Humiles</i> and <i>Proceri</i>	188
9.5.2	Character Evolution and Biogeography	189
9.5.3	Phylogenetics of Hybrids and Polyploids	189
9.6	Conclusions	190
CHAPTER 10	Comparison of Linked versus Unlinked Character Models for Species Tree Inference	191
10.1	Introduction	191
10.2	Methods	192
10.2.1	Simulations of Error-Free Data Sets	192
10.2.2	Introduction of Site Pattern Errors	193
10.2.3	Assessment of Sensitivity to Errors	194
10.2.4	Project Repository	194
10.3	Results	195
10.3.1	Behavior of Linked (StarBEAST2) versus Unlinked (Ecoevolity) Character Models	195
10.3.2	Analysis of All Sites versus SNPs with Ecoevolity	195
10.3.3	Coverage of Credible Intervals	197
10.3.4	MCMC Convergence and Mixing	197
10.4	Discussion	197
10.4.1	Robustness to Character-Pattern Errors	207
10.4.2	Relevance to Empirical Data Sets	208
10.4.3	Recommendations for Using Unlinked-Character Models	209
10.4.4	Other Complexities of Empirical Data in Need of Exploration	209
PART III	Beyond the Species Tree	211
CHAPTER 11	The Unfinished Synthesis of Comparative Genomics and Phylogenetics: Examples from Flightless Birds	215
11.1	Introduction	215
11.1.1	Phylogenetics of Modern Birds	216
11.1.2	Paleognathous Birds as a Test Case for Post-Genomic Phylogenetics	218
11.2	Building of a Whole-Genome Species Tree for an Ancient Radiation of Birds	218
11.3	The Unfinished Synthesis of Comparative Genomics and Genomic Heterogeneity	225
11.3.1	A Species Tree for Paleognathous Birds as a Foundation for Comparative Genomics	225

11.3.2	Accommodation of Uncertainty into Whole-Genome Alignments	225
11.3.3	Gene Tree Heterogeneity and Detecting Rate Variation in Genes and Noncoding Regions	228
11.3.4	Phylogenetic Analysis of Quantitative 'Omics Data: Gene Expression and Epigenetics	230
11.4	Conclusions	231
CHAPTER 12	Phylogenetic Analysis under Heterogeneity and Discordance	232
12.1	Introduction	232
12.2	The Origin of Discordance	232
12.2.1	A History of Systems and Methods	232
12.2.2	Concepts of Harmony and Discordance	234
12.2.3	The Species Tree	236
12.2.4	Comparison of the Incomparable	238
12.3	Characterization and Quantification of Phylogenetic Heterogeneity	238
12.3.1	Quantification and Visualization of Discordance	238
12.3.2	Quantification of Conflict and Tree Evaluation	240
12.3.3	Visualization of Conflict	241
12.4	Analysis under Phylogenetic Heterogeneity	243
12.4.1	Testing of Introgression and Hybridization under Phylogenetic Heterogeneity	243
12.4.2	Testing of Selection under Phylogenetic Heterogeneity	245
12.4.3	Testing of Traits under Phylogenetic Heterogeneity	247
12.4.4	Testing of Coevolution under Phylogenetic Heterogeneity	249
12.5	Conclusion	250
CHAPTER 13	The Multispecies Coalescent in Space and Time	251
13.1	Introduction	251
13.2	Coalescent Simulations	252
13.2.1	Units, Space, and Time	253
13.2.2	Tree Size, Tree Space, and Phylogenetic Decay	255
13.3	Linked Genealogies and Gene Tree Inference	256
13.4	Conclusions	258
CHAPTER 14	Tree Set Visualization, Exploration, and Applications	260
14.1	Introduction to Visualizing and Exploring Tree Sets	260
14.1.1	Tree Set Visualization	261
14.1.2	Detection of Structure in Tree Sets	262
14.2	Applications to Gene Trees, Species Trees, and Phylogenomics	264
14.2.1	Sensitivity to Models of Sequence Evolution	264
14.2.2	Joint versus Independent Inference of Gene Trees	268

14.2.3 Understanding of Variation across Genomes	271
14.2.4 Prospects for Future Development and Application	275
14.3 Appendix	275
<i>Bibliography</i>	277
<i>Index</i>	317