

Contents

List of Figures	xiii
List of Tables	xxi
Preface	xxv
Editors	xxvii
Contributors	xxix
1 Bayesian phylogenetics: methods, computational algorithms, and applications	1
1.1 Introduction	1
1.2 Overview	2
2 Priors in Bayesian phylogenetics	5
<i>Ying Wang and Ziheng Yang</i>	
2.1 Introduction	6
2.1.1 The Bayesian paradigm	6
2.1.2 Objective versus subjective priors	6
2.2 Estimation of distance between two sequences	8
2.2.1 The maximum likelihood estimate (MLE)	8
2.2.2 Uniform or flat priors	8
2.2.3 The Jeffreys priors	9
2.2.4 Reference priors	10
2.3 Priors on model parameters in Bayesian phylogenetics	11
2.3.1 Priors on branch lengths	11
2.3.2 Priors on parameters in substitution models	15
2.3.3 Priors for heterogeneous substitution rates among sites and over time	17
2.3.4 Partition and mixture models for analysis of large genomic datasets	18
2.4 Priors on the tree topology	19
2.4.1 Prior on rooted trees	19
2.4.2 Priors on unrooted trees and impossibility of equal clade probabilities	20

2.4.3	Prior on trees and high posterior probabilities	21
2.5	Priors on times and rates for estimation of divergence times	22
2.6	Summary	23
3	IDR for marginal likelihood in Bayesian phylogenetics	25
	<i>Serena Arima and Luca Tardella</i>	
3.1	Introduction	25
3.2	Substitution models: a brief overview	26
3.2.1	Bayesian inference for substitution models	27
3.3	Bayesian model choice	28
3.4	Computational tools for Bayesian model evidence	32
3.4.1	Harmonic mean estimators	32
3.4.2	IDR: inflated density ratio estimator	34
3.4.2.1	IDR: numerical examples	38
3.4.2.2	IDR for substitution models	41
3.4.3	Thermodynamic integration estimators	43
3.4.3.1	Generalized stepping-stone estimator (GSS)	45
3.4.3.2	Comparative performance: the linear model	47
3.5	Marginal likelihood for phylogenetic data	48
3.5.1	Hadamard data: marginal likelihood computation	49
3.5.2	Green plant <i>rbcL</i> example	53
3.6	Discussion	55
4	Bayesian model selection in phylogenetics and genealogy-based population genetics	59
	<i>Guy Baele and Philippe Lemey</i>	
4.1	Introduction	60
4.2	Prior and posterior-based estimators	62
4.2.1	Integrating the likelihood against the model prior	62
4.2.2	The arithmetic mean estimator	62
4.2.3	The harmonic mean estimator(s)	63
4.2.4	Akaike and Bayesian Information Criterion through Markov chain Monte Carlo: AICM and BICM	64
4.2.5	Generalized harmonic mean estimator (GHME) and inflated density ratio (IDR)	65
4.2.5.1	GHME	65
4.2.5.2	Inflated density ratio (IDR)	66
4.3	Path sampling approaches	67
4.3.1	Path sampling	67
4.3.1.1	Marginal likelihood estimation	69
4.3.1.2	Direct Bayes factor estimation	69
4.3.2	Stepping-stone sampling	70
4.3.2.1	Marginal likelihood estimation	71
4.3.2.2	Direct Bayes factor estimation	72
4.3.3	Generalized stepping-stone sampling	74

4.4	Simulation study: uncorrelated relaxed clocks	76
4.4.1	Relaxed molecular clock models	76
4.4.2	Comparing model selection to model-averaging	78
4.4.3	Simulation settings	81
4.4.4	Results	82
4.5	Practical example on demographic models	85
4.5.1	HIV-1 group M dataset	85
4.5.2	Demographic/coalescent models	86
4.5.3	Proper priors and the Bayesian skyride model	87
4.5.4	Computational requirements	89
4.5.5	Results	90
5	Variable tree topology stepping-stone marginal likelihood estimation	95
	<i>Mark T. Holder, Paul O. Lewis, David L. Swofford, and David Bryant</i>	
5.1	Introduction	96
5.2	The generalized stepping-stone (GSS) method	96
5.3	Reference distribution for tree topology	98
5.3.1	Tree topology reference distribution	98
5.3.1.1	Tree simulation	99
5.3.1.2	Tree topology reference distribution	100
5.3.2	Edge length reference distribution	103
5.3.3	Comparison with CCD methods	104
5.4	Example	105
5.4.1	Model details	105
5.4.2	Brute-force approach	106
5.4.3	GSS performance	108
5.5	Summary	110
5.6	Funding	110
6	Consistency of marginal likelihood estimation when topology varies	113
	<i>Rui Wu, Ming-Hui Chen, Lynn Kuo, and Paul O. Lewis</i>	
6.1	Introduction	114
6.2	Notation and definitions	114
6.2.1	Consistency of VHM	115
6.2.2	Consistency of VSS	116
6.2.3	Consistency of VIDR	119
6.3	Empirical example	120
6.3.1	Transformations used for VIDR	122
6.3.2	Topology-specific marginal likelihood estimation	122
6.3.3	Total marginal likelihood estimation	123
6.4	Discussion	125
6.5	Funding	126

7 Bayesian phylogeny analysis	129
<i>Sooyoung Cheon and Faming Liang</i>	
7.1 Introduction	129
7.2 Bayesian phylogeny inference	130
7.2.1 Tree representation	130
7.2.2 Evolutionary models	132
7.2.3 Bayesian approach	134
7.3 Monte Carlo methods for Bayesian phylogeny inference	135
7.3.1 Markov chain Monte Carlo algorithms	135
7.3.1.1 Rannala and Yang (1996)	136
7.3.1.2 Mau and Newton (1997)	137
7.3.1.3 Mau et al. (1999) and Newton et al. (1999)	139
7.3.1.4 Larget and Simon (1999)	140
7.3.1.5 Li et al. (2000)	142
7.3.2 Advanced MCMC algorithms	144
7.3.2.1 Huelsenbeck and Ronquist (2001)	144
7.3.2.2 Feng et al. (2003)	146
7.3.2.3 Altekari et al. (2004)	149
7.3.2.4 Stochastic approximation Monte Carlo	151
7.3.2.5 Sequential stochastic approximation Monte Carlo	154
7.4 Summary	162
8 SMC (sequential Monte Carlo) for Bayesian phylogenetics	163
<i>Alexandre Bouchard-Côté</i>	
8.1 Using phylogenetic SMC samplers	165
8.1.1 Overview	165
8.1.2 A general framework for understanding the output of Monte Carlo algorithms	168
8.2 How phylogenetic SMC works	169
8.2.1 The foundations: importance sampling	170
8.2.2 Toward SMC: sequential importance sampling (SIS)	171
8.2.3 Resampling	174
8.2.4 Example: Yule trees	176
8.2.5 Inferring non-clock trees	177
8.3 Extensions and implementation issues	178
8.3.1 Efficiently computing and storing the particles	178
8.3.2 More SMC proposals	180
8.3.3 More on resampling	181
8.3.4 Estimating the marginal likelihood	182
8.3.5 Combining SMC and MCMC	183
8.4 Discussion	184

9 Population model comparison using multi-locus datasets	187
<i>Michal Palczewski and Peter Beerli</i>	
9.1 Introduction	187
9.2 Bayesian inference of independent loci	189
9.2.1 What K represents qualitatively	193
9.2.2 Calculating K	194
9.3 Model comparison using our independent marginal likelihood sampler	196
9.4 Conclusion	199
10 Bayesian methods in the presence of recombination	201
<i>Mary K. Kuhner</i>	
10.1 Introduction to non-treelike phylogenies	201
10.2 Describing the ARG	204
10.3 Inference of the ARG	211
10.4 Mechanics of sampling ARGs	216
10.5 Hazards of Bayesian inference in the presence of recombination	221
10.6 Directions for future research	224
10.7 Open questions	226
10.8 Conclusions	227
11 Bayesian nonparametric phylodynamics	229
<i>Julia A. Palacios, Mandev S. Gll, Marc A. Suchard, and Vladimir N. Minin</i>	
11.1 Introduction	230
11.2 General model formulation	232
11.2.1 Likelihood for sequence alignment	232
11.2.2 Coalescent prior	233
11.2.3 Posterior inference	235
11.3 Priors on effective population size trajectory	236
11.3.1 Multiple change-point models	236
11.3.2 Coalescent as a point process	237
11.3.3 Gaussian process-based nonparametrics	237
11.4 Examples	239
11.4.1 Fixed genealogy	239
11.4.2 Accounting for genealogical uncertainty	241
11.5 Extensions and future directions	242
11.5.1 Multiple loci	243
11.5.2 Effect of population structure	244
11.5.3 Coalescent and infectious disease dynamics	245
12 Sampling and summary statistics of endpoint-conditioned paths in DNA sequence evolution	247
<i>Asger Hobolth and Jeffrey L. Thorne</i>	
12.1 Introduction	248

12.2	Independent sites models and summary statistics	250
12.2.1	Likelihood inference	250
12.2.2	The EM algorithm	251
12.2.3	Bayesian inference	252
12.2.4	Conditional means on a phylogeny	252
12.2.5	Endpoint-conditioned summary statistics from uniformization	253
12.3	Dependent-site models and Markov chain Monte Carlo	260
12.3.1	Gibbs sampling with context dependence	262
12.3.2	Path sampling	264
12.3.2.1	Rejection sampling	265
12.3.2.2	Direct sampling	266
12.3.2.3	Uniformization	268
12.3.2.4	Bisectioning	270
12.3.3	Metropolis-Hastings algorithm with dependence	271
12.4	Future directions for sequence paths with dependence models	273
13	Bayesian inference of species divergence times	277
	<i>Tracy A. Heath and Brian R. Moore</i>	
13.1	Introduction	277
13.2	Priors on branch rates	282
13.2.1	Autocorrelated-rate models	282
13.2.2	Independent-rate models	286
13.2.3	Local molecular clock models	289
13.2.4	Mixture models on branch rates	291
13.3	Priors on node times	295
13.3.1	Generic priors on node ages	295
13.3.2	Branching-process priors on node ages	296
13.4	Priors for calibrating divergence times	299
13.4.1	Hierarchical models for calibrating node ages	303
13.4.1.1	Evaluating hierarchical calibration priors	304
13.5	Practical issues for estimating divergence times	308
13.5.1	Model selection and model uncertainty	308
13.5.2	(Hyper)prior sensitivity	311
13.5.3	Limitations and challenges of fossil calibration	312
13.6	Summary and prospectus	316
	Bibliography	319
	Index	359