

Contents

1	Models of nucleotide substitution	1
1.1	Introduction	1
1.2	Markov models of nucleotide substitution and distance estimation	4
1.2.1	The JC69 model	4
1.2.2	The K80 model	7
1.2.3	HKY85, F84, TN93, etc.	9
1.2.4	The transition/transversion rate ratio	13
1.3	Variable substitution rates across sites	15
1.4	Maximum likelihood estimation of distance	17
1.4.1	The JC69 model	18
1.4.2	The K80 model	22
1.4.3	Likelihood ratio test of substitution models	22
*1.4.4	Profile and integrated likelihood methods	24
1.5	Markov chains and distance estimation under general models	26
1.5.1	Markov chains	26
*1.5.2	Distance under the unrestricted (UNREST) model	27
*1.5.3	Distance under the general time-reversible model	29
1.6	Discussions	32
1.6.1	Distance estimation under different substitution models	32
1.6.2	Limitations of pairwise comparison	32
1.7	Problems	33
2	Models of amino acid and codon substitution	35
2.1	Introduction	35
2.2	Models of amino acid replacement	35
2.2.1	Empirical models	35
2.2.2	Mechanistic models	39
2.2.3	Among-site heterogeneity	39
2.3	Estimation of distance between two protein sequences	40
2.3.1	The Poisson model	40
2.3.2	Empirical models	41
2.3.3	Gamma distances	41
2.4	Models of codon substitution	42
2.4.1	The basic model	42
2.4.2	Variations and extensions	44
2.5	Estimation of d_S and d_N	47
2.5.1	Counting methods	47
2.5.2	Maximum likelihood method	55

2.5.3	Comparison of methods	57
2.5.4	More distances and interpretation of the d_N/d_S ratio	58
2.5.5	Estimation of d_S and d_N in comparative genomics	61
*2.5.6	Distances based on the physical-site definition	63
*2.5.7	Utility of the distance measures	65
*2.6	Numerical calculation of the transition probability matrix	65
2.7	Problems	68
3	Phylogeny reconstruction: overview	70
3.1	Tree concepts	70
3.1.1	Terminology	70
3.1.2	Species trees and gene trees	79
3.1.3	Classification of tree reconstruction methods	81
3.2	Exhaustive and heuristic tree search	82
3.2.1	Exhaustive tree search	82
3.2.2	Heuristic tree search	82
3.2.3	Branch swapping	84
3.2.4	Local peaks in the tree space	86
3.2.5	Stochastic tree search	88
3.3	Distance matrix methods	88
3.3.1	Least-squares method	89
3.3.2	Minimum evolution method	91
3.3.3	Neighbour-joining method	91
3.4	Maximum parsimony	95
3.4.1	Brief history	95
3.4.2	Counting the minimum number of changes on a tree	95
3.4.3	Weighted parsimony and dynamic programming	96
3.4.4	Probabilities of ancestral states	99
3.4.5	Long-branch attraction	99
3.4.6	Assumptions of parsimony	100
3.5	Problems	101
4	Maximum likelihood methods	102
4.1	Introduction	102
4.2	Likelihood calculation on tree	102
4.2.1	Data, model, tree, and likelihood	102
4.2.2	The pruning algorithm	103
4.2.3	Time reversibility, the root of the tree, and the molecular clock	107
4.2.4	A numerical example: phylogeny of apes	108
4.2.5	Amino acid, codon, and RNA models	110
*4.2.6	Missing data, sequence errors, and alignment gaps	110
4.3	Likelihood calculation under more complex models	114
4.3.1	Mixture models for variable rates among sites	114
4.3.2	Mixture models for pattern heterogeneity among sites	122
4.3.3	Partition models for combined analysis of multiple datasets	123
4.3.4	Nonhomogeneous and nonstationary models	125

4.4	Reconstruction of ancestral states	125
4.4.1	Overview	125
4.4.2	Empirical and hierarchical Bayesian reconstruction	127
*4.4.3	Discrete morphological characters	130
4.4.4	Systematic biases in ancestral reconstruction	131
*4.5	Numerical algorithms for maximum likelihood estimation	133
*4.5.1	Univariate optimization	134
*4.5.2	Multivariate optimization	136
4.6	ML optimization in phylogenetics	138
4.6.1	Optimization on a fixed tree	138
4.6.2	Multiple local peaks on the likelihood surface for a fixed tree	139
4.6.3	Search in the tree space	140
4.6.4	Approximate likelihood method	143
4.7	Model selection and robustness	144
4.7.1	Likelihood ratio test applied to rbcL dataset	144
4.7.2	Test of goodness of fit and parametric bootstrap	146
*4.7.3	Diagnostic tests to detect model violations	147
4.7.4	Akaike information criterion (AIC and AIC _c)	148
4.7.5	Bayesian information criterion	149
4.7.6	Model adequacy and robustness	150
4.8	Problems	151
5	Comparison of phylogenetic methods and tests on trees	153
5.1	Statistical performance of tree reconstruction methods	153
5.1.1	Criteria	154
5.1.2	Performance	156
5.2	Likelihood	157
5.2.1	Contrast with conventional parameter estimation	157
5.2.2	Consistency	158
5.2.3	Efficiency	159
5.2.4	Robustness	163
5.3	Parsimony	165
5.3.1	Equivalence with misbehaved likelihood models	165
5.3.2	Equivalence with well-behaved likelihood models	168
5.3.3	Assumptions and justifications	169
5.4	Testing hypotheses concerning trees	171
5.4.1	Bootstrap	172
5.4.2	Interior-branch test	177
5.4.3	K-H test and related tests	178
5.4.4	Example: phylogeny of apes	179
5.4.5	Indexes used in parsimony analysis	180
5.5	Problems	181
6	Bayesian theory	182
6.1	Overview	182
6.2	The Bayesian paradigm	183

6.2.1	The Bayes theorem	183
6.2.2	The Bayes theorem in Bayesian statistics	184
*6.2.3	Classical versus Bayesian statistics	189
6.3	Prior	197
6.3.1	Methods of prior specification	197
6.3.2	Conjugate priors	198
6.3.3	Flat or uniform priors	199
*6.3.4	The Jeffreys priors	200
*6.3.5	The reference priors	202
6.4	Methods of integration	203
*6.4.1	Laplace approximation	203
6.4.2	Mid-point and trapezoid methods	204
6.4.3	Gaussian quadrature	205
6.4.4	Marginal likelihood calculation for JC69 distance estimation	206
6.4.5	Monte Carlo integration	210
6.4.6	Importance sampling	210
6.5	Problems	212
7	Bayesian computation (MCMC)	214
7.1	Markov chain Monte Carlo	214
7.1.1	Metropolis algorithm	214
7.1.2	Asymmetrical moves and proposal ratio	218
7.1.3	The transition kernel	219
7.1.4	Single-component Metropolis–Hastings algorithm	220
7.1.5	Gibbs sampler	221
7.2	Simple moves and their proposal ratios	221
7.2.1	Sliding window using the uniform proposal	222
7.2.2	Sliding window using the normal proposal	223
7.2.3	Bactrian proposal	223
7.2.4	Sliding window using the multivariate normal proposal	224
7.2.5	Proportional scaling	225
7.2.6	Proportional scaling with bounds	226
7.3	Convergence, mixing, and summary of MCMC	226
7.3.1	Convergence and tail behaviour	226
7.3.2	Mixing efficiency, jump probability, and step length	230
7.3.3	Validating and diagnosing MCMC algorithms	241
7.3.4	Potential scale reduction statistic	242
7.3.5	Summary of MCMC output	243
7.4	Advanced Monte Carlo methods	244
7.4.1	Parallel tempering (MC ³)	245
7.4.2	Trans-model and trans-dimensional MCMC	247
7.4.3	Bayes factor and marginal likelihood	256
7.5	Problems	260
8	Bayesian phylogenetics	263
8.1	Overview	263
8.1.1	Historical background	263

8.1.2	A sketch MCMC algorithm	264
8.1.3	The statistical nature of phylogeny estimation	264
8.2	Models and priors in Bayesian phylogenetics	266
8.2.1	Priors on branch lengths	266
8.2.2	Priors on parameters in substitution models	269
8.2.3	Priors on tree topology	276
8.3	MCMC proposals in Bayesian phylogenetics	279
8.3.1	Within-tree moves	279
8.3.2	Cross-tree moves	281
8.3.3	NNI for unrooted trees	284
8.3.4	SPR for unrooted trees	287
8.3.5	TBR for unrooted trees	289
8.3.6	Subtree swapping	291
8.3.7	NNI for rooted trees	292
8.3.8	SPR on rooted trees	293
8.3.9	Node slider	294
8.4	Summarizing MCMC output	295
8.5	High posterior probabilities for trees	296
8.5.1	High posterior probabilities for trees or splits	296
8.5.2	Star tree paradox	298
*8.5.3	Fair coin paradox, fair balance paradox, and Bayesian model selection	300
8.5.4	Conservative Bayesian phylogenetics	305
8.6	Problems	306
9	Coalescent theory and species trees	308
9.1	Overview	308
9.2	The coalescent model for a single species	309
9.2.1	The backward time machine	309
9.2.2	Fisher–Wright model and the neutral coalescent	309
9.2.3	A sample of n genes	312
9.2.4	Simulating the coalescent	315
9.2.5	Estimation of θ from a sample of DNA sequences	316
9.3	Population demographic process	320
9.3.1	Homogeneous and nonhomogeneous Poisson processes	321
9.3.2	Deterministic population size change	322
9.3.3	Nonparametric population demographic models	323
9.4	Multispecies coalescent, species trees and gene trees	325
9.4.1	Multispecies coalescent	325
9.4.2	Species tree–gene tree conflict	331
9.4.3	Estimation of species trees	335
9.4.4	Migration	343
9.5	Species delimitation	349
9.5.1	Species concept and species delimitation	349
9.5.2	Simple methods for analysing genetic data	351
9.5.3	Bayesian species delimitation	352

9.5.4	The impact of guide tree, prior, and migration	355
9.5.5	Pros and cons of Bayesian species delimitation	358
9.6	Problems	359
10	Molecular clock and estimation of species divergence times	361
10.1	Overview	361
10.2	Tests of the molecular clock	363
10.2.1	Relative-rate tests	363
10.2.2	Likelihood ratio test	364
10.2.3	Limitations of molecular clock tests	365
10.2.4	Index of dispersion	366
10.3	Likelihood estimation of divergence times	366
10.3.1	Global clock model	366
10.3.2	Local clock model	367
10.3.3	Heuristic rate-smoothing methods	368
10.3.4	Uncertainties in calibrations	370
10.3.5	Dating viral divergences	372
10.3.6	Dating primate divergences	373
10.4	Bayesian estimation of divergence times	375
10.4.1	General framework	375
10.4.2	Approximate calculation of likelihood	376
10.4.3	Prior on evolutionary rates	377
10.4.4	Prior on divergence times and fossil calibrations	378
10.4.5	Uncertainties in time estimates	382
10.4.6	Dating viral divergences	384
10.4.7	Application to primate and mammalian divergences	385
10.5	Perspectives	388
10.6	Problems	389
11	Neutral and adaptive protein evolution	390
11.1	Introduction	390
11.2	The neutral theory and tests of neutrality	391
11.2.1	The neutral and nearly neutral theories	391
11.2.2	Tajima's D statistic	393
11.2.3	Fu and Li's D , and Fay and Wu's H statistics	394
11.2.4	McDonald–Kreitman test and estimation of selective strength	395
11.2.5	Hudson–Kreitman–Aquade test	397
11.3	Lineages undergoing adaptive evolution	398
11.3.1	Heuristic methods	398
11.3.2	Likelihood method	399
11.4	Amino acid sites undergoing adaptive evolution	400
11.4.1	Three strategies	400
11.4.2	Likelihood ratio test of positive selection under random-site models	402
11.4.3	Identification of sites under positive selection	405
11.4.4	Positive selection at the human MHC	406

11.5	Adaptive evolution affecting particular sites and lineages	408
11.5.1	Branch-site test of positive selection	408
11.5.2	Other similar models	409
11.5.3	Adaptive evolution in angiosperm phytochromes	410
11.6	Assumptions, limitations, and comparisons	411
11.6.1	Assumptions and limitations of current methods	412
11.6.2	Comparison of methods for detecting positive selection	413
11.7	Adaptively evolving genes	414
11.8	Problems	416
12	Simulating molecular evolution	418
12.1	Introduction	418
12.2	Random number generator	418
12.3	Generation of discrete random variables	420
12.3.1	Inversion method for sampling from a general discrete distribution	420
12.3.2	The alias method for sampling from a discrete distribution	421
12.3.3	Discrete uniform distribution	422
12.3.4	Binomial distribution	423
12.3.5	The multinomial distribution	423
12.3.6	The Poisson distribution	423
12.3.7	The composition method for mixture distributions	424
12.4	Generation of continuous random variables	424
12.4.1	The inversion method	425
12.4.2	The transformation method	425
12.4.3	The rejection method	425
12.4.4	Generation of a standard normal variate using the polar method	428
12.4.5	Gamma, beta, and Dirichlet variables	430
12.5	Simulation of Markov processes	430
12.5.1	Simulation of the Poisson process	430
12.5.2	Simulation of the nonhomogeneous Poisson process	431
12.5.3	Simulation of discrete-time Markov chains	433
12.5.4	Simulation of continuous-time Markov chains	435
12.6	Simulating molecular evolution	436
12.6.1	Simulation of sequences on a fixed tree	436
12.6.2	Simulation of random trees	439
12.7	Validation of the simulation program	439
12.8	Problems	440
	Appendices	442
	Appendix A. Functions of random variables	442
	Appendix B. The delta technique	446
	Appendix C. Phylogenetic software	448
	<i>References</i>	450
	<i>Index</i>	488