

Contents

	PART I: MODELLING MOLECULAR EVOLUTION	1
I	Models of nucleotide substitution	3
1.1	Introduction	3
1.2	Markov models of nucleotide substitution and distance estimation	4
1.2.1	The JC69 model	4
1.2.2	The K80 model	10
1.2.3	HKY85, F84, TN93, etc.	11
1.2.4	The transition/transversion rate ratio	17
1.3	Variable substitution rates across sites	18
1.4	Maximum likelihood estimation	22
1.4.1	The JC69 model	22
1.4.2	The K80 model	25
*1.4.3	Profile and integrated likelihood methods	27
1.5	Markov chains and distance estimation under general models	30
1.5.1	General theory	30
1.5.2	The general time-reversible (GTR) model	33
1.6	Discussions	37
1.6.1	Distance estimation under different substitution models	37
1.6.2	Limitations of pairwise comparison	37
1.7	Exercises	38
2	Models of amino acid and codon substitution	40
2.1	Introduction	40
2.2	Models of amino acid replacement	40
2.2.1	Empirical models	40
2.2.2	Mechanistic models	43
2.2.3	Among-site heterogeneity	44
2.3	Estimation of distance between two protein sequences	45
2.3.1	The Poisson model	45

2.3.2	Empirical models	46
2.3.3	Gamma distances	46
2.3.4	Example: distance between cat and rabbit p53 genes	47
2.4	Models of codon substitution	48
2.5	Estimation of synonymous and nonsynonymous substitution rates	49
2.5.1	Counting methods	50
2.5.2	Maximum likelihood method	58
2.5.3	Comparison of methods	61
*2.5.4	Interpretation and a plethora of distances	62
*2.6	Numerical calculation of the transition-probability matrix	68
2.7	Exercises	70

PART II: PHYLOGENY RECONSTRUCTION

3	Phylogeny reconstruction: overview	71
3.1	Tree concepts	73
3.1.1	Terminology	73
3.1.2	Topological distance between trees	77
3.1.3	Consensus trees	79
3.1.4	Gene trees and species trees	80
3.1.5	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	83
3.2.3	Branch swapping	84
3.2.4	Local peaks in the tree space	87
3.2.5	Stochastic tree search	89
3.3	Distance methods	89
3.3.1	Least-squares method	90
3.3.2	Neighbour-joining method	92
3.4	Maximum parsimony	93
3.4.1	Brief history	93
3.4.2	Counting the minimum number of changes given the tree	94
3.4.3	Weighted parsimony and transversion parsimony	95
3.4.4	Long-branch attraction	98
3.4.5	Assumptions of parsimony	99
4	Maximum likelihood methods	100
4.1	Introduction	100
4.2	Likelihood calculation on tree	100
4.2.1	Data, model, tree, and likelihood	100
4.2.2	The pruning algorithm	102
4.2.3	Time reversibility, the root of the tree and the molecular clock	106

4.2.4	Missing data and alignment gaps	107
4.2.5	An numerical example: phylogeny of apes	108
4.3	Likelihood calculation under more complex models	109
4.3.1	Models of variable rates among sites	110
4.3.2	Models for combined analysis of multiple data sets	116
4.3.3	Nonhomogeneous and nonstationary models	118
4.3.4	Amino acid and codon models	119
4.4	Reconstruction of ancestral states	119
4.4.1	Overview	119
4.4.2	Empirical and hierarchical Bayes reconstruction	121
*4.4.3	Discrete morphological characters	124
4.4.4	Systematic biases in ancestral reconstruction	126
*4.5	Numerical algorithms for maximum likelihood estimation	128
4.5.1	Univariate optimization	129
4.5.2	Multivariate optimization	131
4.5.3	Optimization on a fixed tree	134
4.5.4	Multiple local peaks on the likelihood surface for a fixed tree	135
4.5.5	Search for the maximum likelihood tree	136
4.6	Approximations to likelihood	137
4.7	Model selection and robustness	137
4.7.1	LRT, AIC, and BIC	137
4.7.2	Model adequacy and robustness	142
4.8	Exercises	144
5	Bayesian methods	145
5.1	The Bayesian paradigm	145
5.1.1	Overview	145
5.1.2	Bayes's theorem	146
5.1.3	Classical versus Bayesian statistics	151
5.2	Prior	158
5.3	Markov chain Monte Carlo	159
5.3.1	Monte Carlo integration	160
5.3.2	Metropolis–Hastings algorithm	161
5.3.3	Single-component Metropolis–Hastings algorithm	164
5.3.4	Gibbs sampler	166
5.3.5	Metropolis-coupled MCMC (MCMCMC or MC ³)	166
5.4	Simple moves and their proposal ratios	167
5.4.1	Sliding window using the uniform proposal	168
5.4.2	Sliding window using normal proposal	168
5.4.3	Sliding window using the multivariate normal proposal	169
5.4.4	Proportional shrinking and expanding	170
5.5	Monitoring Markov chains and processing output	171
5.5.1	Validating and diagnosing MCMC algorithms	171

5.5.2	Potential scale reduction statistic	173
5.5.3	Processing output	174
5.6	Bayesian phylogenetics	174
5.6.1	Brief history	174
5.6.2	General framework	175
5.6.3	Summarizing MCMC output	175
5.6.4	Bayesian versus likelihood	177
5.6.5	A numerical example: phylogeny of apes	180
5.7	MCMC algorithms under the coalescent model	181
5.7.1	Overview	181
5.7.2	Estimation of θ	181
5.8	Exercises	184
6	Comparison of methods and tests on trees	185
6.1	Statistical performance of tree-reconstruction methods	186
6.1.1	Criteria	186
6.1.2	Performance	188
6.2	Likelihood	190
6.2.1	Contrast with conventional parameter estimation	190
6.2.2	Consistency	191
6.2.3	Efficiency	192
6.2.4	Robustness	196
6.3	Parsimony	198
6.3.1	Equivalence with misbehaved likelihood models	198
6.3.2	Equivalence with well-behaved likelihood models	201
6.3.3	Assumptions and Justifications	204
6.4	Testing hypotheses concerning trees	206
6.4.1	Bootstrap	207
6.4.2	Interior branch test	210
6.4.3	Kishino-Hasegawa test and modifications	211
6.4.4	Indexes used in parsimony analysis	213
6.4.5	Example: phylogeny of apes	214
*6.5	Appendix: Tuffley and Steel's likelihood analysis of one character	215
	PART III: ADVANCED TOPICS	221
7	Molecular clock and estimation of species divergence times	223
7.1	Overview	223
7.2	Tests of the molecular clock	225
7.2.1	Relative-rate tests	225
7.2.2	Likelihood ratio test	226

7.2.3	Limitations of the clock tests	227
7.2.4	Index of dispersion	228
7.3	Likelihood estimation of divergence times	228
7.3.1	Global-clock model	228
7.3.2	Local-clock models	230
7.3.3	Heuristic rate-smoothing methods	231
7.3.4	Dating primate divergences	233
*7.3.5	Uncertainties in fossils	235
7.4	Bayesian estimation of divergence times	245
7.4.1	General framework	245
7.4.2	Calculation of the likelihood	246
7.4.3	Prior on rates	247
7.4.4	Uncertainties in fossils and prior on divergence times	248
7.4.5	Application to primate and mammalian divergences	252
7.5	Perspectives	257
8	Neutral and adaptive protein evolution	259
8.1	Introduction	259
8.2	The neutral theory and tests of neutrality	260
8.2.1	The neutral and nearly neutral theory	260
8.2.2	Tajima's D statistic	262
8.2.3	Fu and Li's D and Fay and Wu's H statistics	264
8.2.4	McDonald-Kreitman test and estimation of selective strength	265
8.2.5	Hudson-Kreitman-Aquade test	267
8.3	Lineages undergoing adaptive evolution	268
8.3.1	Heuristic methods	268
8.3.2	Likelihood method	269
8.4	Amino acid sites undergoing adaptive evolution	271
8.4.1	Three strategies	271
8.4.2	Likelihood ratio test of positive selection under random-sites models	273
8.4.3	Identification of sites under positive selection	276
8.4.4	Positive selection in the human major histocompatibility (MHC) locus	276
8.5	Adaptive evolution affecting particular sites and lineages	279
8.5.1	Branch-site test of positive selection	279
8.5.2	Other similar models	281
8.5.3	Adaptive evolution in angiosperm phytochromes	282
8.6	Assumptions, limitations, and comparisons	284
8.6.1	Limitations of current methods	284
8.6.2	Comparison between tests of neutrality and tests based on d_N and d_S	286
8.7	Adaptively evolving genes	286

9	Simulating molecular evolution	293
9.1	Introduction	293
9.2	Random number generator	294
9.3	Generation of continuous random variables	295
9.4	Generation of discrete random variables	296
9.4.1	Discrete uniform distribution	296
9.4.2	Binomial distribution	297
9.4.3	General discrete distribution	297
9.4.4	Multinomial distribution	298
9.4.5	The composition method for mixture distributions	298
*9.4.6	The alias method for sampling from a discrete distribution	299
9.5	Simulating molecular evolution	302
9.5.1	Simulating sequences on a fixed tree	302
9.5.2	Generating random trees	305
9.6	Exercises	306
10	Perspectives	308
10.1	Theoretical issues in phylogeny reconstruction	308
10.2	Computational issues in analysis of large and heterogeneous data sets	309
10.3	Genome rearrangement data	309
10.4	Comparative genomics	310
Appendices		311
A:	Functions of random variables	311
B:	The delta technique	313
C:	Phylogenetics software	316
References		319
Index		353