

Contents

Preface	xix
1 Parsimony methods	1
A simple example	1
Evaluating a particular tree	1
Rootedness and unrootedness	4
Methods of rooting the tree	6
Branch lengths	8
Unresolved questions	9
2 Counting evolutionary changes	11
The Fitch algorithm	11
The Sankoff algorithm	13
Connection between the two algorithms	16
Using the algorithms when modifying trees	16
Views	16
Using views when a tree is altered	17
Further economies	18
3 How many trees are there?	19
Rooted bifurcating trees	20
Unrooted bifurcating trees	24
Multifurcating trees	25
Unrooted trees with multifurcations	28
Tree shapes	29
Rooted bifurcating tree shapes	29
Rooted multifurcating tree shapes	30
Unrooted Shapes	32
Labeled histories	35
Perspective	36

4	Finding the best tree	
	by heuristic search	37
	Nearest-neighbor interchanges	38
	Subtree pruning and regrafting	41
	Tree bisection and reconnection	44
	Other tree rearrangement methods	44
	Tree-fusing	44
	Genetic algorithms	45
	Tree windows and sectorial search	46
	Speeding up rearrangements	46
	Sequential addition	47
	Star decomposition	48
	Tree space	50
	Search by reweighting of characters	51
	Simulated annealing	52
	History	53
5	Finding the best tree	
	by branch and bound	54
	A nonbiological example	54
	Finding the optimal solution	57
	NP-hardness	59
	Branch and bound methods	60
	Phylogenies: Despair and hope	60
	Branch and bound for parsimony	61
	Improving the bound	64
	Using still-absent states	64
	Using compatibility	64
	Rules limiting the search	65
6	Ancestral states	
	and branch lengths	67
	Reconstructing ancestral states	67
	Accelerated and delayed transformation	70
	Branch lengths	70
7	Variants of parsimony	73
	Camin-Sokal parsimony	73
	Parsimony on an ordinal scale	74
	Dollo parsimony	75
	Polymorphism parsimony	76
	Unknown ancestral states	78
	Multiple states and binary coding	78
	Dollo parsimony and multiple states	80

Polymorphism parsimony and multiple states	81
Transformation series analysis	81
Weighting characters	82
Successive weighting and nonlinear weighting	83
Successive weighting	83
Nonsuccessive algorithms	85
8 Compatibility	87
Testing compatibility	88
The Pairwise Compatibility Theorem	89
Cliques of compatible characters	91
Finding the tree from the clique	92
Other cases where cliques can be used	94
Where cliques cannot be used	94
Perfect phylogeny	95
Using compatibility on molecules anyway	96
9 Statistical properties of parsimony	97
Likelihood and parsimony	97
The weights	100
Unweighted parsimony	100
Limitations of this justification of parsimony	101
Farris's proofs	102
No common mechanism	103
Likelihood and compatibility	105
Parsimony versus compatibility	107
Consistency and parsimony	107
Character patterns and parsimony	107
Observed numbers of the patterns	110
Observed fractions of the patterns	111
Expected fractions of the patterns	111
Inconsistency	113
When inconsistency is not a problem	114
The nucleotide sequence case	115
Other situations where consistency is guaranteed	117
Does a molecular clock guarantee consistency?	118
The Farris zone	120
Some perspective	121
10 A digression on history and philosophy	123
How phylogeny algorithms developed	123
Sokal and Sneath	123
Edwards and Cavalli-Sforza	125
Camin and Sokal and parsimony	129

Eck and Dayhoff and molecular parsimony	130
Fitch and Margoliash popularize distance matrix methods	131
Wilson and Le Quesne introduce compatibility	133
Jukes and Cantor and molecular distances	134
Farris and Kluge and unordered parsimony	134
Fitch and molecular parsimony	136
Further work	136
What about Willi Hennig and Walter Zimmerman?	136
Different philosophical frameworks	138
Hypothetico-deductive	139
Logical parsimony	141
Logical probability?	142
Criticisms of statistical inference	143
The irrelevance of classification	145
11 Distance matrix methods	147
Branch lengths and times	147
The least squares methods	148
Least squares branch lengths	148
Finding the least squares tree topology	153
The statistical rationale	153
Generalized least squares	154
Distances	155
The Jukes-Cantor model—an example	156
Why correct for multiple changes?	158
Minimum evolution	159
Clustering algorithms	161
UPGMA and least squares	161
A clustering algorithm	162
An example	162
UPGMA on nonclocklike trees	165
Neighbor-joining	166
Performance	168
Using neighbor-joining with other methods	169
Relation of neighbor-joining to least squares	169
Weighted versions of neighbor-joining	170
Other approximate distance methods	171
Distance Wagner method	171
A related family	171
Minimizing the maximum discrepancy	172
Two approaches to error in trees	172
A puzzling formula	173
Consistency and distance methods	174

A limitation of distance methods	175
12 Quartets of species	176
The four point metric	177
The split decomposition	178
Related methods	182
Short quartets methods	182
The disk-covering method	183
Challenges for the short quartets and DCM methods	185
Three-taxon statement methods	186
Other uses of quartets with parsimony	188
Consensus supertrees	189
Neighborliness	191
De Soete's search method	192
Quartet puzzling and searching tree space	193
Perspective	194
13 Models of DNA evolution	196
Kimura's two-parameter model	196
Calculation of the distance	198
The Tamura-Nei model, F84, and HKY	200
The general time-reversible model	204
Distances from the GTR model	206
The general 12-parameter model	210
LogDet distances	211
Other distances	213
Variance of distance	214
Rate variation between sites or loci	215
Different rates at different sites	215
Distances with known rates	216
Distribution of rates	216
Gamma- and lognormally distributed rates	217
Distances from gamma-distributed rates	217
Models with nonindependence of sites	221
14 Models of protein evolution	222
Amino acid models	222
The Dayhoff model	222
Other empirically-based models	223
Models depending on secondary structure	225
Codon-based models	226
Inequality of synonymous and nonsynonymous substitutions	227
Protein structure and correlated change	228

15 Restriction sites, RAPDs, AFLPs, and microsatellites	230
Restriction sites	230
Nei and Tajima's model	230
Distances based on restriction sites	233
Issues of ascertainment	234
Parsimony for restriction sites	235
Modeling restriction fragments	236
Parsimony with restriction fragments	239
RAPDs and AFLPs	239
The issue of dominance	240
Unresolved problems	240
Microsatellite models	241
The one-step model	241
Microsatellite distances	242
A Brownian motion approximation	244
Models with constraints on array size	246
Multi-step and heterogeneous models	246
Snakes and Ladders	246
Complications	247
16 Likelihood methods	248
Maximum likelihood	248
An example	249
Computing the likelihood of a tree	251
Economizing on the computation	253
Handling ambiguity and error	255
Unrootedness	256
Finding the maximum likelihood tree	256
Inferring ancestral sequences	259
Rates varying among sites	260
Hidden Markov models	262
Autocorrelation of rates	264
HMMs for other aspects of models	265
Estimating the states	265
Models with clocks	266
Relaxing molecular clocks	266
Models for relaxed clocks	267
Covarions	268
Empirical approaches to change of rates	269
Are ML estimates consistent?	269
Comparability of likelihoods	270
A nonexistent proof?	270
A simple proof	271

Misbehavior with the wrong model	272
Better behavior with the wrong model	274
17 Hadamard methods	275
The edge length spectrum and conjugate spectrum	279
The closest tree criterion	281
DNA models	284
Computational effort	285
Extensions of Hadamard methods	286
18 Bayesian inference of phylogenies	288
Bayes' theorem	288
Bayesian methods for phylogenies	291
Markov chain Monte Carlo methods	292
The Metropolis algorithm	292
Its equilibrium distribution	293
Bayesian MCMC	294
Bayesian MCMC for phylogenies	295
Priors	295
Proposal distributions	296
Computing the likelihoods	298
Summarizing the posterior	299
Priors on trees	300
Controversies over Bayesian inference	301
Universality of the prior	301
Flat priors and doubts about them	302
Applications of Bayesian methods	304
19 Testing models, trees, and clocks	307
Likelihood and tests	307
Likelihood ratios near asymptopia	308
Multiple parameters	309
Some parameters constrained, some not	310
Conditions	310
Curvature or height?	311
Interval estimates	311
Testing assertions about parameters	312
Coins in a barrel	313
Evolutionary rates instead of coins	314
Choosing among nonnested hypotheses: AIC and BIC	315
An example using the AIC criterion	317
The problem of multiple topologies	318
LRTs and single branches	319
Interior branch tests	320

Interior branch tests using parsimony	321
A multiple-branch counterpart of interior branch tests	322
Testing the molecular clock	322
Parsimony-based methods	323
Distance-based methods	323
Likelihood-based methods	323
The relative rate test	324
Simulation tests based on likelihood	328
Further literature	329
More exact tests and confidence intervals	329
Tests for three species with a clock	330
Bremer support	330
Zander's conditional probability of reconstruction	331
More generalized confidence sets	332
20 Bootstrap, jackknife, and permutation tests	335
The bootstrap and the jackknife	335
Bootstrapping and phylogenies	337
The delete-half jackknife	339
The bootstrap and jackknife for phylogenies	340
The multiple-tests problem	342
Independence of characters	343
Identical distribution — a problem?	343
Invariant characters and resampling methods	344
Biases in bootstrap and jackknife probabilities	346
<i>P</i> values in a simple normal case	346
Methods of reducing the bias	349
The drug testing analogy	352
Alternatives to <i>P</i> values	355
Probabilities of trees	356
Using tree distances	356
Jackknifing species	357
Parametric bootstrapping	357
Advantages and disadvantages of the parametric bootstrap	358
Permutation tests	359
Permuting species within characters	359
Permuting characters	361
Skewness of tree length distribution	362
21 Paired-sites tests	364
An example	365
Multiple trees	369
The SH test	370
Other multiple-comparison tests	371

Testing other parameters	372
Perspective	372
22 Invariants	373
Symmetry invariants	375
Three-species invariants	376
Lake's linear invariants	378
Cavender's quadratic invariants	380
The K invariants	380
The L invariants	381
Generalization of Cavender's L invariants	382
Drolet and Sankoff's k -state quadratic invariants	385
Clock invariants	385
General methods for finding invariants	386
Fourier transform methods	386
Gröbner bases and other general methods	387
Expressions for all the 3ST invariants	387
Finding all invariants empirically	388
All linear invariants	388
Special cases and extensions	389
Invariants and evolutionary rates	389
Testing invariants	389
What use are invariants?	390
23 Brownian motion and gene frequencies	391
Brownian motion	391
Likelihood for a phylogeny	392
What likelihood to compute?	395
Assuming a clock	399
The REML approach	401
Multiple characters and Kronecker products	402
Pruning the likelihood	404
Maximizing the likelihood	406
Inferring ancestral states	408
Squared-change parsimony	409
Gene frequencies and Brownian motion	410
Using approximate Brownian motion	412
Distances from gene frequencies	412
A more exact likelihood method	413
Gene frequency parsimony	413

24 Quantitative characters	415
Neutral models of quantitative characters	416
Changes due to natural selection	419
Selective correlation	419
Covariances of multiple characters in multiple lineages	420
Selection for an optimum	421
Brownian motion and selection	422
Correcting for correlations	422
Punctuational models	424
Inferring phylogenies and correlations	425
Chasing a common optimum	426
The character-coding "problem"	426
Continuous-character parsimony methods	428
Manhattan metric parsimony	428
Other parsimony methods	429
Threshold models	429
25 Comparative methods	432
An example with discrete states	432
An example with continuous characters	433
The contrasts method	435
Correlations between characters	436
When the tree is not completely known	437
Inferring change in a branch	438
Sampling error	439
The standard regression and other variations	442
Generalized least squares	442
Phylogenetic autocorrelation	442
Transformations of time	442
Should we use the phylogeny at all?	443
Paired-lineage tests	443
Discrete characters	444
Ridley's method	444
Concentrated-changes tests	445
A paired-lineages test	446
Methods using likelihood	446
Advantages of the likelihood approach	448
Molecular applications	448
26 Coalescent trees	450
Kingman's coalescent	454
Bugs in a box—an analogy	460
Effect of varying population size	460
Migration	461

Effect of recombination	464
Coalescents and natural selection	467
Neuhauser and Krone's method	468
27 Likelihood calculations on coalescents	470
The basic equation	470
Using accurate genealogies—a reverie	471
Two random sampling methods	473
A Metropolis-Hastings method	473
Griffiths and Tavaré's method	476
Bayesian methods	482
MCMC for a variety of coalescent models	482
Single-tree methods	484
Slatkin and Maddison's method	484
Fu's method	485
Summary-statistic methods	485
Watterson's method	485
Other summary-statistic methods	486
Testing for recombination	486
28 Coalescents and species trees	488
Methods of inferring the species phylogeny	491
Reconciled tree parsimony approaches	492
Likelihood	493
29 Alignment, gene families, and genomics	496
Alignment	497
Why phylogenies are important	497
Parsimony method	498
Approximations and progressive alignment	500
Probabilistic models	502
Bishop and Thompson's method	502
The minimum message length method	502
The TKF model	503
Multibase insertions and deletions	506
Tree HMMs	507
Trees	507
Inferring the alignment	509
Gene families	509
Reconciled trees	510
Reconstructing duplications	511
Rooting unrooted trees	512
A likelihood analysis	514
Comparative genomics	515

Tandemly repeated genes	515
Inversions	516
Inversions in trees	516
Inversions, transpositions, and translocations	516
Breakpoint and neighbor-coding approximations	517
Synten	517
Probabilistic models	518
Genome signature methods	519
30 Consensus trees and distances between trees	521
Consensus trees	521
Strict consensus	522
Majority-rule consensus	523
Adams consensus tree	524
A dismaying result	525
Consensus using branch lengths	526
Other consensus tree methods	526
Consensus subtrees	528
Distances between trees	528
The symmetric difference	529
The quartets distance	530
The nearest-neighbor interchange distance	530
The path-length-difference metric	531
Distances using branch lengths	531
Are these distances truly distances?	533
Consensus trees and distances	534
Trees significantly the same? different?	534
What do consensus trees and tree distances tell us?	535
The total evidence debate	536
A modest proposal	537
31 Biogeography, hosts, and parasites	539
Component compatibility	540
Brooks parsimony	541
Event-based parsimony methods	543
Relation to tree reconciliation	545
Randomization tests	545
Statistical inference	546
32 Phylogenies and paleontology	547
Stratigraphic indices	548
Stratophenetics	549
Stratocladistics	550
Controversies	552

A not-quite-likelihood method	553
Stratolikelihood	553
Making a full likelihood method	554
More realistic fossilization models	554
Fossils within species: Sequential sampling	555
Between species	557
33 Tests based on tree shape	559
Using the topology only	559
Imbalance at the root	560
Harding's probabilities of tree shapes	561
Tests from shapes	562
Measures of overall asymmetry	563
Choosing a powerful test	564
Tests using times	564
Lineage plots	565
Likelihood formulas	567
Other likelihood approaches	569
Other statistical approaches	569
A time transformation	570
Characters and key innovations	571
Work remaining	571
34 Drawing trees	573
Issues in drawing rooted trees	574
Placement of interior nodes	574
Shapes of lineages	576
Unrooted trees	578
The equal-angle algorithm	578
n-Body algorithms	580
The equal-daylight algorithm	582
Challenges	584
35 Phylogeny software	585
Trees, records, and pointers	585
Declaring records	586
Traversing the tree	587
Unrooted tree data structures	589
Tree file formats	590
Widely used phylogeny programs and packages	591
References	595
Index	645