

The Phylogenetic Handbook

**A Practical Approach to Phylogenetic
Analysis and Hypothesis Testing**

Second Edition

Edited by

Philippe Lemey

Katholieke Universiteit Leuven, Belgium

Marco Salemi

University of Florida, Gainesville, USA

Anne-Mieke Vandamme

Katholieke Universiteit Leuven, Belgium

Contents

List of contributors

page xix

Foreward

xxiii

Preface

xxv

Section I: Introduction

1

Basic concepts of molecular evolution

3

Anne-Mieke Vandamme

1.1 Genetic information

3

1.2 Population dynamics

9

1.3 Evolution and speciation

14

1.4 Data used for molecular phylogenetics

16

1.5 What is a phylogenetic tree?

19

1.6 Methods for inferring phylogenetic trees

23

1.7 Is evolution always tree-like?

28

Section II: Data preparation

31

2 Sequence databases and database searching

33

Theory

33

Guy Bottu

2.1 Introduction

33

2.2 Sequence databases

35

2.2.1 General nucleic acid sequence databases

35

2.2.2 General protein sequence databases

37

2.2.3 Specialized sequence databases, reference databases, and

2.3.2	Sequence Retrieval System (SRS)	43
2.3.3	Some general considerations about database searching by keyword	44
2.4	Database searching by sequence similarity	45
2.4.1	Optimal alignment	45
2.4.2	Basic Local Alignment Search Tool (BLAST)	47
2.4.3	FASTA	50
2.4.4	Other tools and some general considerations	52
	Practice	55
	Marc Van Ranst and Philippe Lemey	
2.5	Database searching using ENTREZ	55
2.6	BLAST	62
2.7	FASTA	66

3 Multiple sequence alignment 68

	Theory	68
	Des Higgins and Philippe Lemey	
3.1	Introduction	68
3.2	The problem of repeats	68
3.3	The problem of substitutions	70
3.4	The problem of gaps	72
3.5	Pairwise sequence alignment	74
3.5.1	Dot-matrix sequence comparison	74
3.5.2	Dynamic programming	75
3.6	Multiple alignment algorithms	79
3.6.1	Progressive alignment	80
3.6.2	Consistency-based scoring	89
3.6.3	Iterative refinement methods	90
3.6.4	Genetic algorithms	90
3.6.5	Hidden Markov models	91
3.6.6	Other algorithms	91
3.7	Testing multiple alignment methods	92
3.8	Which program to choose?	93
3.9	Nucleotide sequences vs. amino acid sequences	95
3.10	Visualizing alignments and manual editing	96

Practice 100

	Des Higgins and Philippe Lemey	
3.11	CUSTAL alignment	100

3.12	T-COFFEE alignment	102
3.13	MUSCLE alignment	102
3.14	Comparing alignments using the ALTAvisT web tool	103
3.15	From protein to nucleotide alignment	104
3.16	Editing and viewing multiple alignments	105
3.17	Databases of alignments	106

Section III: Phylogenetic inference 109

4 Genetic distances and nucleotide substitution models 111

Theory 111

Korbinian Strimmer and Arndt von Haeseler

4.1	Introduction	111
4.2	Observed and expected distances	112
4.3	Number of mutations in a given time interval *(<i>optional</i>)	113
4.4	Nucleotide substitutions as a <i>homogeneous Markov process</i>	116
4.4.1	The Jukes and Cantor (JC69) model	117
4.5	Derivation of Markov Process *(<i>optional</i>)	118
4.5.1	Inferring the expected distances	121
4.6	Nucleotide substitution models	121
4.6.1	Rate heterogeneity among sites	123

Practice 126

Marco Salemi

4.7	Software packages	126
4.8	Observed vs. estimated genetic distances: the JC69 model	128
4.9	Kimura 2-parameters (K80) and F84 <i>genetic distances</i>	131
4.10	More complex models	132
4.10.1	Modeling rate heterogeneity among sites	133
4.11	Estimating standard errors using MEGA4	135
4.12	The problem of substitution saturation	137
4.13	Choosing among different evolutionary models	140

5 Phylogenetic inference based on distance methods 142

Theory 142

Yves Van de Peer

5.1	Introduction	142
5.2	Tree-inference methods based on genetic distances	144

5.3	Evaluating the reliability of inferred trees	150
5.3.1	Bootstrap analysis	151
5.3.2	Jackknifing	151
5.4	Conclusions	151

Practice

Marco Salemi

5.5	Programs to display and manipulate phylogenetic trees	161
5.6	Distance-based phylogenetic inference in PHYLIP	162
5.7	Inferring a Neighbor-Joining tree for the primates data set	163
5.7.1	Outgroup rooting	168
5.8	Inferring a <i>Fitch–Margoliash</i> tree for the mtDNA data set	170
5.9	Bootstrap analysis using PHYLIP	170
5.10	Impact of genetic distances on tree topology: an example using MEGA4	174
5.11	Other programs	180

6 Phylogenetic inference using maximum likelihood methods

Theory

Heiko A. Schmidt and Arndt von Haeseler

6.1	Introduction	181
6.2	The formal framework	184
6.2.1	The simple case: maximum-likelihood tree for two sequences	184
6.2.2	The complex case	185
6.3	Computing the probability of an alignment for a fixed tree	186
6.3.1	Felsenstein’s pruning algorithm	188
6.4	Finding a maximum-likelihood tree	189
6.4.1	Early heuristics	190
6.4.2	Full-tree rearrangement	190
6.4.3	DNAML and FASTDNAML	191
6.4.4	PHYML and PHYML-SPR	192
6.4.5	IQPNNI	192
6.4.6	RAXML	193
6.4.7	Simulated annealing	193
6.4.8	Genetic algorithms	194
6.5	Branch support	194
6.6	The quartet puzzling algorithm	195
6.6.1	Parameter estimation	195

Practice	199
Heiko A. Schmidt and Arndt von Haeseler	
6.8 Software packages	199
6.9 An illustrative example of an ML tree reconstruction	199
6.9.1 Reconstructing an ML tree with IQPNNI	199
6.9.2 Getting a tree with branch support values using quartet puzzling	203
6.9.3 Likelihood-mapping analysis of the HIV data set	207
6.10 Conclusions	207
7 Bayesian phylogenetic analysis using MrBAYES	210
Theory	210
Fredrik Ronquist, Paul van der Mark, and John P. Huelsenbeck	
7.1 Introduction	210
7.2 Bayesian phylogenetic inference	216
7.3 Markov chain Monte Carlo sampling	220
7.4 Burn-in, mixing and convergence	224
7.5 Metropolis coupling	227
7.6 Summarizing the results	229
7.7 An introduction to phylogenetic models	230
7.8 Bayesian model choice and model averaging	232
7.9 Prior probability distributions	236
Practice	237
Fredrik Ronquist, Paul van der Mark, and John P. Huelsenbeck	
7.10 Introduction to MrBAYES	237
7.10.1 Acquiring and installing the program	237
7.10.2 Getting started	238
7.10.3 Changing the size of the MrBAYES window	238
7.10.4 Getting help	239
7.11 A simple analysis	240
7.11.1 Quick start version	240
7.11.2 Getting data into MrBAYES	241
7.11.3 Specifying a model	242
7.11.4 Setting the priors	244
7.11.5 Checking the model	247
7.11.6 Setting up the analysis	248
7.11.7 Running the analysis	252
7.11.8 When to stop the analysis	254

7.12	Analyzing a partitioned data set	261
7.12.1	Getting mixed data into MRBAYES	261
7.12.2	Dividing the data into partitions	261
7.12.3	Specifying a partitioned model	263
7.12.4	Running the analysis	265
7.12.5	Some practical advice	265

Phylogeny inference based on parsimony and other methods using PAUP* 267

Theory 267

David L. Swofford and Jack Sullivan

8.1	Introduction	267
8.2	Parsimony analysis – background	268
8.3	Parsimony analysis – methodology	270
8.3.1	Calculating the length of a given tree under the parsimony criterion	270
8.4	Searching for optimal trees	273
8.4.1	Exact methods	277
8.4.2	Approximate methods	282

Practice 289

David L. Swofford and Jack Sullivan

8.5	Analyzing data with PAUP* through the command–line interface	292
8.6	Basic parsimony analysis and tree-searching	293
8.7	Analysis using distance methods	300
8.8	Analysis using maximum likelihood methods	303

Phylogenetic analysis using protein sequences 313

Theory 313

Fred R. Opperdoes

9.1	Introduction	313
9.2	Protein evolution	314
9.2.1	Why analyze protein sequences?	314
9.2.2	The genetic code and codon bias	315
9.2.3	Look-back time	317
9.2.4	Nature of sequence divergence in proteins (the PAM unit)	319
9.2.5	Introns and non-coding DNA	321
9.2.6	Choosing DNA or protein?	322

Practice	332
Fred R. Opperdoes and Philippe Lemey	
9.4 A phylogenetic analysis of the Leishmanial glyceraldehyde-3-phosphate dehydrogenase gene carried out via the Internet	332
9.5 A phylogenetic analysis of trypanosomatid glyceraldehyde-3-phosphate dehydrogenase protein sequences using Bayesian inference	337

Section IV: Testing models and trees 343

10 Selecting models of evolution 345

Theory 345

David Posada

10.1 Models of evolution and phylogeny reconstruction	345
10.2 Model fit	346
10.3 Hierarchical likelihood ratio tests (hLRTs)	348
10.3.1 Potential problems with the hLRTs	349
10.4 Information criteria	349
10.5 Bayesian approaches	351
10.6 Performance-based selection	352
10.7 Model selection uncertainty	352
10.8 Model averaging	353

Practice 355

David Posada

10.9 The model selection procedure	355
10.10 MODELTEST	355
10.11 PROTTEST	358
10.12 Selecting the best-fit model in the example data sets	359
10.12.1 Vertebrate mtDNA	359
10.12.2 HIV-1 envelope gene	360
10.12.3 G3PDH protein	361

11 Molecular clock analysis 362

Theory 362

Philippe Lemey and David Posada

11.3	Likelihood ratio test of the global molecular clock	365
11.4	Dated tips	367
11.5	Relaxing the molecular clock	369
11.6	Discussion and future directions	371
	Practice	373
	Philippe Lemey and David Posada	
11.7	Molecular clock analysis using PAML	373
11.8	Analysis of the primate sequences	375
11.9	Analysis of the viral sequences	377
12	Testing tree topologies	381
	Theory	381
	Heiko A. Schmidt	
12.1	Introduction	381
12.2	Some definitions for distributions and testing	382
12.3	Likelihood ratio tests for nested models	384
12.4	How to get the distribution of likelihood ratios	385
	12.4.1 Non-parametric bootstrap	386
	12.4.2 Parametric bootstrap	387
12.5	Testing tree topologies	387
	12.5.1 Tree tests – a general structure	388
	12.5.2 The original Kishino–Hasegawa (KH) test	388
	12.5.3 One-sided Kishino–Hasegawa test	389
	12.5.4 Shimodaira–Hasegawa (SH) test	390
	12.5.5 Weighted test variants	390
	12.5.6 The approximately unbiased test	392
	12.5.7 Swofford–Olsen–Waddell–Hillis (SOWH) test	393
12.6	Confidence sets based on likelihood weights	394
12.7	Conclusions	395
	Practice	397
	Heiko A. Schmidt	
12.8	Software packages	397
12.9	Testing a set of trees with TREE-PUZZLE and CONSEL	397
	12.9.1 Testing and obtaining site-likelihood with TREE-PUZZLE	398
	12.9.2 Testing with CONSEL	401

Section V: Molecular adaptation	405
13 Natural selection and adaptation of molecular sequences	407
Oliver G. Pybus and Beth Shapiro	
13.1 Basic concepts	407
13.2 The molecular footprint of selection	412
13.2.1 Summary statistic methods	413
13.2.2 d_N/d_S methods	415
13.2.3 Codon volatility	417
13.3 Conclusion	418
14 Estimating selection pressures on alignments of coding sequences	419
Theory	419
Sergei L. Kosakovsky Pond, Art F. Y. Poon, and Simon D. W. Frost	
14.1 Introduction	419
14.2 Prerequisites	423
14.3 Codon substitution models	424
14.4 Simulated data: how and why?	426
14.5 Statistical estimation procedures	426
14.5.1 Distance-based approaches	426
14.5.2 Maximum likelihood approaches	428
14.5.3 Estimating d_S and d_N	429
14.5.4 Correcting for nucleotide substitution biases	431
14.5.5 Bayesian approaches	438
14.6 Estimating branch-by-branch variation in rates	438
14.6.1 Local vs. global model	439
14.6.2 Specifying branches <i>a priori</i>	439
14.6.3 Data-driven branch selection	440
14.7 Estimating site-by-site variation in rates	442
14.7.1 Random effects likelihood (REL)	442
14.7.2 Fixed effects likelihood (FEL)	445
14.7.3 Counting methods	446
14.7.4 Which method to use?	447
14.7.5 The importance of synonymous rate variation	449
14.8 Comparing rates at a site in different branches	449
14.9 Discussion and further directions	450
Practice	452
Sergei L. Kosakovsky Pond, Art F. Y. Poon, and Simon D. W. Frost	

14.10.3	MEGA	453
14.10.4	HYPHY	453
14.10.5	DATAMONKEY	454
14.11	Influenza A as a case study	454
14.12	Prerequisites	455
14.12.1	Getting acquainted with HYPHY	455
14.12.2	Importing alignments and trees	456
14.12.3	Previewing sequences in HYPHY	457
14.12.4	Previewing trees in HYPHY	459
14.12.5	Making an alignment	461
14.12.6	Estimating a tree	462
14.12.7	Estimating nucleotide biases	464
14.12.8	Detecting recombination	465
14.13	Estimating global rates	467
14.13.1	Fitting a global model in the HYPHY GUI	467
14.13.2	Fitting a global model with a HYPHY batch file	470
14.14	Estimating branch-by-branch variation in rates	470
14.14.1	Fitting a local codon model in HYPHY	471
14.14.2	Interclade variation in substitution rates	473
14.14.3	Comparing internal and terminal branches	474
14.15	Estimating site-by-site variation in rates	475
14.15.1	Preliminary analysis set-up	476
14.15.2	Estimating β/α	477
14.15.3	Single-likelihood ancestor counting (SLAC)	477
14.15.4	Fixed effects likelihood (FEL)	478
14.15.5	REL methods in HYPHY	481
14.16	Estimating gene-by-gene variation in rates	484
14.16.1	Comparing selection in different populations	484
14.16.2	Comparing selection between different genes	485
14.17	Automating choices for HYPHY analyses	487
14.18	Simulations	488
14.19	Summary of standard analyses	488
14.20	Discussion	490

Section VI: Recombination

15 Introduction to recombination detection

15.3	Linkage disequilibrium, substitution patterns, and evolutionary inference	495
15.4	Evolutionary implications of recombination	496
15.5	Impact on phylogenetic analyses	498
15.6	Recombination analysis as a multifaceted discipline	506
15.6.1	Detecting recombination	506
15.6.2	Recombinant identification and breakpoint detection	507
15.6.3	Recombination rate	507
15.7	Overview of recombination detection tools	509
15.8	Performance of recombination detection tools	517

16 Detecting and characterizing individual recombination events 519

Theory 519

Mika Salminen and Darren Martin

16.1	Introduction	519
16.2	Requirements for detecting recombination	520
16.3	Theoretical basis for recombination detection methods	523
16.4	Identifying and characterizing actual recombination events	530

Practice 532

Mika Salminen and Darren Martin

16.5	Existing tools for recombination analysis	532
16.6	Analyzing example sequences to detect and characterize individual recombination events	533
16.6.1	Exercise 1: Working with SIMPLOT	533
16.6.2	Exercise 2: Mapping recombination with SIMPLOT	536
16.6.3	Exercise 3: Using the “groups” feature of SIMPLOT	537
16.6.4	Exercise 4: Setting up RDP3 to do an exploratory analysis	538
16.6.5	Exercise 5: Doing a simple exploratory analysis with RDP3	540
16.6.6	Exercise 6: Using RDP3 to refine a recombination hypothesis	546

Section VII: Population genetics 549

17 The coalescent: population genetic inference using genealogies 551

Allen Rodrigo

17.1	Introduction	551
------	--------------	-----

17.4	The mutation clock	555
17.5	Demographic history and the coalescent	556
17.6	Coalescent-based inference	558
17.7	The serial coalescent	559
17.8	Advanced topics	561
18	Bayesian evolutionary analysis by sampling trees	564
	Theory	564
	Alexei J. Drummond and Andrew Rambaut	
18.1	Background	564
18.2	Bayesian MCMC for genealogy-based population genetics	566
	18.2.1 Implementation	567
	18.2.2 Input format	568
	18.2.3 Output and results	568
	18.2.4 Computational performance	568
18.3	Results and discussion	569
	18.3.1 Substitution models and rate models among sites	570
	18.3.2 Rate models among branches, divergence time estimation, and time-stamped data	570
	18.3.3 Tree priors	571
	18.3.4 Multiple data partitions and linking and unlinking parameters	572
	18.3.5 Definitions and units of the standard parameters and variables	572
	18.3.6 Model comparison	572
	18.3.7 Conclusions	575
	Practice	576
	Alexei J. Drummond and Andrew Rambaut	
18.4	The BEAST software package	576
18.5	Running BEAUTI	576
18.6	Loading the NEXUS file	577
18.7	Setting the dates of the taxa	577
	18.7.1 Translating the data in amino acid sequences	579
18.8	Setting the evolutionary model	579
18.9	Setting up the operators	580
18.10	Setting the MCMC options	581
18.11	Running BEAST	582
18.12	Analyzing the BEAST output	583
18.13	Summarizing the trees	586

19	LAMARC: Estimating population genetic parameters from molecular data	592
	Theory	592
	Mary K. Kuhner	
	19.1 Introduction	592
	19.2 Basis of the Metropolis–Hastings MCMC sampler	593
	19.2.1 Bayesian vs. likelihood sampling	595
	19.2.2 Random sample	595
	19.2.3 Stability	596
	19.2.4 No other forces	596
	19.2.5 Evolutionary model	596
	19.2.6 Large population relative to sample	597
	19.2.7 Adequate run time	597
	Practice	598
	Mary K. Kuhner	
	19.3 The LAMARC software package	598
	19.3.1 FLUCTUATE (COALESCE)	598
	19.3.2 MIGRATE-N	598
	19.3.3 RECOMBINE	599
	19.3.4 LAMARC	600
	19.4 Starting values	600
	19.5 Space and time	601
	19.6 Sample size considerations	601
	19.7 Virus-specific issues	602
	19.7.1 Multiple loci	602
	19.7.2 Rapid growth rates	603
	19.7.3 Sequential samples	603
	19.8 An exercise with LAMARC	603
	19.8.1 Converting data using the LAMARC file converter	604
	19.8.2 Estimating the population parameters	605
	19.8.3 Analyzing the output	607
	19.9 Conclusions	611
	Section VIII: Additional topics	613
20	Assessing substitution saturation with DAMBE	615
	Theory	615
	Xuhua Xia	

20.3	Xia's method: its problem, limitation, and implementation in DAMBE	621
	Practice	624
	Xuhua Xia and Philippe Lemey	
20.4	Working with the VertebrateMtCOI.FAS file	624
20.5	Working with the InvertebrateEF1a.FAS file	628
20.6	Working with the SIV.FAS file	629
21	Split networks. A tool for exploring complex evolutionary relationships in molecular data	631
	Theory	631
	Vincent Moulton and Katharina T. Huber	
21.1	Understanding evolutionary relationships through networks	631
21.2	An introduction to split decomposition theory	633
	21.2.1 The Buneman tree	634
	21.2.2 Split decomposition	636
21.3	From weakly compatible splits to networks	638
21.4	Alternative ways to compute split networks	639
	21.4.1 NeighborNet	639
	21.4.2 Median networks	640
	21.4.3 Consensus networks and supernetworks	640
	Practice	642
	Vincent Moulton and Katharina T. Huber	
21.5	The SPLITSTREE program	642
	21.5.1 Introduction	642
	21.5.2 Downloading SPLITSTREE	642
21.6	Using SPLITSTREE on the mtDNA data set	642
	21.6.1 Getting started	643
	21.6.2 The fit index	643
	21.6.3 Laying out split networks	645
	21.6.4 Recomputing split networks	645
	21.6.5 Computing trees	646
	21.6.6 Computing different networks	646
	21.6.7 Bootstrapping	646
	21.6.8 Printing	647
21.7	Using SPLITSTREE on other data sets	648