

# The Phylogenetic Handbook

## A Practical Approach to DNA and Protein Phylogeny

Edited by

Marco Salemi

University of California, Irvine and Katholieke Universiteit Leuven, Belgium

and

Anne-Mieke Vandamme

Rega Institute for Medical Research, Katholieke Universiteit Leuven, Belgium

Universitäts- und Landes-  
bibliothek Darmstadt  
Bibliothek Biologie

Inv.-Nr. 15632  
.....  
.....



**CAMBRIDGE**  
UNIVERSITY PRESS

# Contents

<i>Foreword</i>	page xvii
<i>Acknowledgments</i>	xxi
<i>Contributors</i>	xxiii
<b>1 Basic concepts of molecular evolution</b>	<b>1</b>
Anne-Mieke Vandamme	
1.1 Genetic information	1
1.2 Population dynamics	6
1.3 Data used for molecular phylogenetic analysis	10
1.4 What is a phylogenetic tree?	14
1.5 Methods to infer phylogenetic trees	17
1.6 Is evolution always tree-like?	21
<b>2 Sequence databases</b>	<b>24</b>
<b>THEORY</b>	<b>24</b>
Guy Bottu and Marc Van Ranst	
2.1 General nucleic acid sequence databases	24
2.2 General protein sequence databases	26
2.3 Nonredundant sequence databases	27
2.4 Specialized sequence databases	28
2.5 Databases with aligned protein sequences	29
2.6 Database documentation search	30
2.6.1 Text-string searching	30
2.6.2 Searching by index	30
2.7 ENTREZ database	32
2.8 Sequence similarity searching: BLAST	33
<b>PRACTICE</b>	<b>37</b>
Marco Salemi	
2.9 File formats	37

2.10	Three example data sets	40
2.10.1	Preparing input files: HIV/SIV example data set	41
<b>3</b>	<b>Multiple alignment</b>	<b>45</b>
	<b>THEORY</b>	<b>45</b>
	Des Higgins	
3.1	Introduction	45
3.2	The problem of repeats	46
3.3	The problem of substitutions	47
3.4	The problem of gaps	50
3.5	Testing multiple-alignment methods	51
3.6	Multiple-alignment algorithms	52
3.6.1	Dot-matrix sequence comparison	52
3.6.2	Dynamic programming	54
3.6.3	Genetic algorithms	55
3.6.4	Other algorithms	55
3.7	Progressive alignment	55
3.7.1	Clustal	57
3.7.2	T-Coffee	58
3.8	Hidden Markov models	58
3.9	Nucleotide sequences versus amino-acid sequences	59
	<b>PRACTICE</b>	<b>61</b>
	Des Higgins and Marco Salemi	
3.10	Searching for homologous sequences with BioEdit	61
3.11	File formats for Clustal	63
3.12	Access to ClustalW and ClustalX	64
3.13	Aligning the HIV/SIV sequences with ClustalX	64
3.14	Aligning nucleotide sequences in a coding region with DAMBE	66
3.15	Adding sequences to preexisting alignments	67
3.16	Editing and viewing multiple alignments	68
3.17	Databases of alignments	69
<b>4</b>	<b>Nucleotide substitution models</b>	<b>72</b>
	<b>THEORY</b>	<b>72</b>
	Korbinian Strimmer and Arndt von Haeseler	
4.1	Introduction	72
4.2	Observed and expected distances	73
4.3	Number of mutations in a given time interval *(optional)	74
4.4	Nucleotide substitutions as a homogeneous Markov process	77
4.4.1	The Jukes and Cantor (JC69) model	79
4.5	Derivation of Markov process *(optional)	80
4.5.1	Inferring the expected distances	83

4.6 Nucleotide substitution models	83
4.6.1 Rate heterogeneity over sites	85
<b>PRACTICE: The PHYLIP and TREE-PUZZLE software packages</b>	88
Marco Salemi	
4.7 Software packages	88
4.8 Jukes and Cantor (JC69) genetic distances	90
4.9 Kimura 2-parameters (K80) and F84 genetic distances	91
4.10 More complex models	92
4.10.1 Modeling rate heterogeneity over sites	93
4.11 The problem of substitution saturation	95
4.12 Choosing among different evolutionary models	97
<b>5 Phylogeny inference based on distance methods</b>	101
<b>THEORY</b>	101
Yves Van de Peer	
5.1 Introduction	101
5.2 Tree-inferring methods based on genetic distances	103
5.2.1 Cluster analysis (UPGMA and WPGMA)	103
5.2.2 Minimum evolution and neighbor-joining	107
5.2.3 Other distance methods	113
5.3 Evaluating the reliability of inferred trees	115
5.3.1 Bootstrap analysis	115
5.3.2 Jackknifing	118
5.4 Conclusions	118
<b>PRACTICE</b>	120
Marco Salemi	
5.5 The TreeView program	120
5.6 Procedure to estimate distance-based phylogenetic trees with PHYLIP	120
5.7 Inferring an NJ tree for the mtDNA data set	121
5.8 Inferring a Fitch-Margoliash tree for the mtDNA data set	125
5.9 Inferring an NJ tree for the HIV-1 data set	125
5.10 Bootstrap analysis with PHYLIP	126
5.11 Other programs	133
<b>6 Phylogeny inference based on maximum-likelihood methods with TREE-PUZZLE</b>	137
<b>THEORY</b>	137
Arndt von Haeseler and Korbinian Strimmer	
6.1 Introduction	137

6.2 The formal framework	140
6.2.1 The simple case: Maximum-likelihood tree for two sequences	140
6.2.2 The complex case	141
6.3 Computing the probability of an alignment for a fixed tree	142
6.3.1 Felsenstein's pruning algorithm	144
6.4 Finding a maximum-likelihood tree	145
6.4.1 The quartet-puzzling algorithm	146
6.5 Estimating the model parameters with maximum likelihood	149
6.6 Likelihood-mapping analysis	150
<b>PRACTICE</b>	153
Arndt von Haeseler and Korbinian Strimmer	
6.7 Software packages	153
6.8 An illustrative example of quartet-puzzling tree reconstruction	153
6.9 Likelihood-mapping analysis of the HIV data set	156

**7 Phylogeny inference based on parsimony and other methods using PAUP\*** 160

---

<b>THEORY</b>	160
David L. Swofford and Jack Sullivan	
6.1 Introduction	160
6.2 Parsimony analysis – background	161
6.3 Parsimony analysis – methodology	163
6.3.1 Calculating the length of a given tree under the parsimony criterion	163
6.4 Searching for optimal trees	166
6.4.1 Exact methods	171
6.4.2 Approximate methods	175
<b>PRACTICE</b>	182
David L. Swofford and Jack Sullivan	
6.5 Analyzing data with PAUP* through the command-line interface	182
6.6 Basic parsimony analysis and tree-searching	186
6.7 Analysis using distance methods	193
6.8 Analysis using maximum-likelihood methods	196

**8 Phylogenetic analysis using protein sequences** 207

---

<b>THEORY</b>	207
Fred R. Opperdoes	
8.1 Introduction	207
8.2 Why protein sequences?	209
8.2.1 The genetic code	210

8.2.2 Codon bias	210
8.2.3 Long time horizon	210
8.2.4 Phylogenetic noise reduction	211
8.2.5 Introns and noncoding DNA	211
8.2.6 Multigene families and post-transcriptional editing	212
8.3 Measurement of sequence divergence in proteins: The PAM	213
8.4 Alignment of protein sequences	215
8.4.1 Sequence retrieval and multiple-sequence alignment	219
8.4.2 Secondary-structure-based alignment	219
8.4.3 Prodom, Pfam, and Blocks databases	220
8.4.4 Manual adjustment of a protein alignment	220
8.5 Tree-building methods for protein phylogeny	221
8.6 Some good advice	224
<b>PRACTICE</b>	226
Fred R. Opperdoes	
8.7 A phylogenetic analysis of the Leismanial GPD gene carried out via the Internet	226
8.8 A comparison of the trypanosomatid phylogeny from nucleotide and protein sequences	230
8.9 Implementing different evolutionary models with DAMBE and TREE-PUZZLE	233
<b>9 Analysis of nucleotide sequences using TREECON</b>	236
<b>THEORY</b>	236
Yves Van de Peer	
9.1 Introduction	236
9.2 TREECON, distance trees, and among-site rate variation	236
9.2.1 Taking into account among-site rate variation: An example	241
9.3 Conclusions	245
<b>PRACTICE</b>	246
Yves Van de Peer	
9.4 The TREECON software package	246
9.5 Implementation	246
9.6 Substitution rate calibration	251
<b>10 Selecting models of evolution</b>	256
<b>THEORY</b>	256
David Posada	
10.1 Models of evolution and phylogeny reconstruction	256
10.2 The relevance of models of evolution	257

10.3	Selecting models of evolution	257
10.4	The likelihood ratio test	258
10.4.1	LRTs and parametric bootstrapping	259
10.4.2	Hierarchical LRTs	260
10.4.3	Dynamical LRTs	261
10.5	Information criteria	263
10.5.1	AIC	264
10.5.2	BIC	264
10.6	Fit of a single model to the data	264
10.7	Testing the molecular clock hypothesis	265
10.7.1	The relative rate test	266
10.7.2	LRT of the global molecular clock	267
	<b>PRACTICE</b>	270
	David Posada	
10.8	The model-selection procedure	270
10.9	The program MODELTEST	273
10.10	Implementing the LRT of the molecular clock using PAUP*	275
10.11	Selecting the best-fit model in the example data sets	276
10.11.1	Vertebrate mtDNA	277
10.11.2	HIV envelope gene	278
10.11.3	G3PDH protein	279
<b>11</b>	<b>Analysis of coding sequences</b>	283
	<b>THEORY</b>	283
	Yoshiyuki Suzuki and Takashi Gojobori	
11.1	Introduction	283
11.2	Mutation fraction methods	285
11.2.1	Method of Nei and Gojobori (NG86 method)	285
11.2.2	Method of Zhang et al. (ZRN98 method)	287
11.2.3	Method of Ina (I95 method)	288
11.3	Degenerate site methods	290
11.3.1	Method of Li et al. (LWL85 method)	291
11.3.2	Method of Pamilo and Bianchi, and Li (PBL93 method)	294
11.4	Codon model methods	294
11.4.1	Method of Muse (M96 method)	295
11.4.2	Method of Yang and Nielsen (YN98 method)	296
11.5	Methods for estimating $d_S$ and $d_N$ at single codon sites	296
11.5.1	Method of Suzuki and Gojobori (SG99 method)	297
11.6	Test of neutrality for two sequences	298
11.6.1	Z test	298
11.6.2	Likelihood ratio test (LRT)	298
11.6.3	Window analysis	299

11.7 Test of neutrality at single codon sites	299
11.7.1 Method of Nielsen and Yang (1998) (NY98 method)	300
11.7.2 SG99 method	300
<b>PRACTICE</b>	302
<i>Yoshiyuki Suzuki and Takashi Gojobori</i>	
11.8 Software for analyzing coding sequences	302
11.9 Estimation of $d_S$ and $d_N$ in an HCV data set	302
11.9.1 Estimation of $d_S$ and $d_N$ with NG86, ZRN98, LWL85, and PBL93 methods (MEGA2)	303
11.9.2 Estimation of $d_S$ and $d_N$ with YN98 method (PAML)	304
11.9.3 Comparing different estimates of $d_S$ and $d_N$	305
11.10 An example of window analysis	306
11.11 Detection of positive selection at single amino acid sites	307
11.12 Conclusions	308
<b>12 SplitsTree: A network-based tool for exploring evolutionary relationships in molecular data</b>	312
<hr/>	
<b>THEORY</b>	312
Vincent Moulton	
12.1 Exploring evolutionary relationships through networks	312
12.2 An introduction to split-decomposition theory	314
12.2.1 The Buneman tree	315
12.2.2 Split decomposition	316
12.3 From weakly compatible splits to networks	318
<b>PRACTICE</b>	320
Vincent Moulton	
12.4 The SplitsTree program	320
12.5 Using SplitsTree on the mtDNA data set	320
12.6 Using SplitsTree on the HIV-1 data set	324
<b>13 Tetrapod phylogeny and data exploration using DAMBE</b>	329
<hr/>	
<b>THEORY</b>	329
Xuhua Xia and Zheng Xie	
13.1 The phylogenetic problem and the sequence data	329
13.2 Results of routine phylogenetic analyses without data exploration	330
13.3 Distance-based statistical test of alternative phylogenetic trees (optional)	332
13.4 Likelihood-based statistical tests of alternative phylogenetic trees	333



13.5	Data exploration	335
13.5.1	Nucleotide frequencies	335
13.5.2	Substitution saturation and the rate heterogeneity over sites	337
13.5.3	The pattern of nucleotide substitution	338
13.5.4	Insertion and deletion as phylogenetic characters	339
	<b>PRACTICE</b>	342
	Xuhua Xia and Zheng Xie	
13.6	Data exploration with DAMBE	342
13.6.1	Nucleotide frequencies	342
13.6.2	Basic phylogenetic reconstruction	342
13.6.3	Rate heterogeneity over sites estimated through reconstruction of ancestral sequences	343
13.6.4	Empirical substitution pattern	344
13.6.5	Testing alternative phylogenetic hypotheses with the distance-based method	344
13.6.6	Testing alternative phylogenetic hypotheses with the likelihood-based method	345
<b>14</b>	<b>Detecting recombination in viral sequences</b>	348
	<b>THEORY</b>	348
	Mika Salminen	
14.1	Introduction and theoretical background to exploring recombination in viral sequences	348
14.2	Requirements for detecting recombination	349
14.3	Theoretical basis for methods to detect recombination	351
14.4	Examples of viral recombination	360
	<b>PRACTICE</b>	362
	Mika Salminen	
14.5	Existing tools for analysis of recombination	362
14.6	Analyzing example sequences to visualize recombination	364
14.6.1	Exercise 1: Working with <code>Simplot</code>	364
14.6.2	Exercise 2: Mapping recombination with <code>Simplot</code>	368
14.6.3	Exercise 3: Using the “groups” feature of <code>Simplot</code>	369
14.6.4	Exercise 4: Using <code>SplitsTree</code> to visualize recombination	373
<b>15</b>	<b>LAMARC: Estimating population genetic parameters from molecular data</b>	378
	<b>THEORY</b>	378
	Mary K. Kuhner	
15.1	Introduction	378

15.2 Basis of the Metropolis-Hastings MCMC sampler	379
15.2.1 Random sample	381
15.2.2 Stability	381
15.2.3 No other forces	381
15.2.4 Evolutionary model	381
15.2.5 Large population relative to sample	382
15.2.6 Adequate run time	382
<b>PRACTICE</b>	384
Mary K. Kuhner	
15.3 The LAMARC software package	384
15.3.1 FLUCTUATE (COALESCE)	384
15.3.2 MIGRATE	384
15.3.3 RECOMBINE	385
15.3.4 LAMARC	386
15.4 Starting values	386
15.5 Space and time	387
15.6 Sample size considerations	387
15.7 Virus-specific issues	388
15.7.1 Multiple loci	388
15.7.2 Rapid growth rates	388
15.7.3 Sequential samples	389
15.8 An exercise with LAMARC	389
15.8.1 Exercise using FLUCTUATE	390
15.8.2 Exercise using RECOMBINE	395
15.9 Conclusions	396
<i>Index</i>	399
<i>Color section follows p. 328.</i>	