The Phylogenetic Handbook
Second Edition

The Phylogenetic Handbook provides a comprehensive introduction to theory and practice of nucleotide and protein phylogenetic analysis. This second edition includes seven new chapters, covering topics such as Bayesian inference, tree topology testing, and the impact of recombination on phylogenies. The book has a stronger focus on hypothesis testing than the previous edition, with more extensive discussions on recombination analysis, detecting molecular adaptation and genealogy-based population genetics. Many chapters include elaborate practical sections, which have been updated to introduce the reader to the most recent versions of sequence analysis and phylogeny software, including BLAST, FASTA, CLUSTAL, T-COFFEE, MUSCLE, DAMBE, TREE-PUZZLE, PHYLIP, MEGA4, PAUP*, IQPNNI, CONSEL, MODELTEST, PROTTEST, PAML, HYPHY, MRBAYES, BEAST, LAMARC, SPLITSTREE, and RDP3. Many analysis tools are described by their original authors, resulting in clear explanations that constitute an ideal teaching guide for advanced-level undergraduate and graduate students.

Philippe Lemey is a FWO postdoctoral researcher at the Rega Institute, Katholieke Universiteit Leuven, Belgium, where he completed his Ph.D. in Medical Sciences. He has been an EMBO Fellow and a Marie-Curie Fellow in the Evolutionary Biology Group at the Department of Zoology, University of Oxford. His research focuses on molecular evolution of viruses by integrating molecular biology and computational approaches.

Marco Salemi is Assistant Professor at the Department of Pathology, Immunology and Laboratory Medicine of the University of Florida School of Medicine, Gainesville, USA. His research interests include molecular epidemiology, intra-host virus evolution, and the application of phylogenetic and population genetic methods to the study of human and simian pathogenic viruses.

Anne-Mieke Vandamme is a Full Professor in the Medical Faculty at the Katholieke Universiteit, Belgium, working in the field of clinical and epidemiological virology. Her laboratory investigates treatment responses in HIV-infected patients and is respected for its scientific and clinical contributions to virus–drug resistance. Her laboratory also studies the evolution and molecular epidemiology of human viruses such as HIV and HTLV.
Contents

List of contributors xix
Foreword 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

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 genome databases 39
2.3 Composite databases, database mirroring, and search tools 39
  2.3.1 Entrez 39
## Contents

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

#### Theory

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

Des Higgins and Philippe Lemey

3.11 CLUSTAL alignment 100
3.11.1 File formats and availability 100
3.11.2 Aligning the primate Trim5α amino acid sequences 101
### Contents

- 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**

### 4 Genetic distances and nucleotide substitution models

#### Theory
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 *(optional)*  
  113
- 4.4 Nucleotide substitutions as a *homogeneous Markov process*  
  116
  - 4.4.1 The Jukes and Cantor (JC69) model  
    117
- 4.5 Derivation of Markov Process *(optional)*  
  118
  - 4.5.1 Inferring the expected distances  
    121
- 4.6 Nucleotide substitution models  
  121
  - 4.6.1 Rate heterogeneity among sites  
    123

#### Practice
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 genetic distances  
  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

#### Theory
Yves Van de Peer

- 5.1 Introduction  
  142
- 5.2 Tree-inference methods based on genetic distances  
  144
  - 5.2.1 Cluster analysis (UPGMA and WPGMA)  
    144
  - 5.2.2 Minimum evolution and neighbor-joining  
    148
  - 5.2.3 Other distance methods  
    156
5.3 Evaluating the reliability of inferred trees 156
  5.3.1 Bootstrap analysis 157
  5.3.2 Jackknifing 159
5.4 Conclusions 159

Practice 161
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 Fitch–Margoliash 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 181

Theory 181
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
  6.6.2 ML step 196
  6.6.3 Puzzling step 196
  6.6.4 Consensus step 196
6.7 Likelihood-mapping analysis 196
ix

Contents

Practice

Heiko A. Schmidt and Arndt von Haeseler

6.8  Software packages
6.9  An illustrative example of an ML tree reconstruction
  6.9.1  Reconstructing an ML tree with IQPNNI
  6.9.2  Getting a tree with branch support values using quartet puzzling
  6.9.3  Likelihood-mapping analysis of the HIV data set
6.10  Conclusions

7  Bayesian phylogenetic analysis using MrBayes

Theory

Fredrik Ronquist, Paul van der Mark, and John P. Huelsenbeck

7.1  Introduction
7.2  Bayesian phylogenetic inference
7.3  Markov chain Monte Carlo sampling
7.4  Burn-in, mixing and convergence
7.5  Metropolis coupling
7.6  Summarizing the results
7.7  An introduction to phylogenetic models
7.8  Bayesian model choice and model averaging
7.9  Prior probability distributions

Practice

Fredrik Ronquist, Paul van der Mark, and John P. Huelsenbeck

7.10  Introduction to MrBayes
  7.10.1  Acquiring and installing the program
  7.10.2  Getting started
  7.10.3  Changing the size of the MrBayes window
  7.10.4  Getting help
7.11  A simple analysis
    7.11.1  Quick start version
    7.11.2  Getting data into MrBayes
    7.11.3  Specifying a model
    7.11.4  Setting the priors
    7.11.5  Checking the model
    7.11.6  Setting up the analysis
    7.11.7  Running the analysis
    7.11.8  When to stop the analysis
    7.11.9  Summarizing samples of substitution model parameters
    7.11.10  Summarizing samples of trees and branch lengths
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.12</td>
<td>Analyzing a partitioned data set</td>
<td>261</td>
</tr>
<tr>
<td>7.12.1</td>
<td>Getting mixed data into MrBayes</td>
<td>261</td>
</tr>
<tr>
<td>7.12.2</td>
<td>Dividing the data into partitions</td>
<td>261</td>
</tr>
<tr>
<td>7.12.3</td>
<td>Specifying a partitioned model</td>
<td>263</td>
</tr>
<tr>
<td>7.12.4</td>
<td>Running the analysis</td>
<td>265</td>
</tr>
<tr>
<td>7.12.5</td>
<td>Some practical advice</td>
<td>265</td>
</tr>
<tr>
<td>8</td>
<td>Phylogeny inference based on parsimony and other methods using Paup*</td>
<td>267</td>
</tr>
<tr>
<td>8.1</td>
<td>Introduction</td>
<td>267</td>
</tr>
<tr>
<td>8.2</td>
<td>Parsimony analysis – background</td>
<td>268</td>
</tr>
<tr>
<td>8.3</td>
<td>Parsimony analysis – methodology</td>
<td>270</td>
</tr>
<tr>
<td>8.3.1</td>
<td>Calculating the length of a given tree under the parsimony criterion</td>
<td>270</td>
</tr>
<tr>
<td>8.4</td>
<td>Searching for optimal trees</td>
<td>273</td>
</tr>
<tr>
<td>8.4.1</td>
<td>Exact methods</td>
<td>277</td>
</tr>
<tr>
<td>8.4.2</td>
<td>Approximate methods</td>
<td>282</td>
</tr>
<tr>
<td>8.5</td>
<td>Analyzing data with Paup* through the command-line interface</td>
<td>292</td>
</tr>
<tr>
<td>8.6</td>
<td>Basic parsimony analysis and tree-searching</td>
<td>293</td>
</tr>
<tr>
<td>8.7</td>
<td>Analysis using distance methods</td>
<td>300</td>
</tr>
<tr>
<td>8.8</td>
<td>Analysis using maximum likelihood methods</td>
<td>303</td>
</tr>
<tr>
<td>9</td>
<td>Phylogenetic analysis using protein sequences</td>
<td>313</td>
</tr>
<tr>
<td>9.1</td>
<td>Introduction</td>
<td>313</td>
</tr>
<tr>
<td>9.2</td>
<td>Protein evolution</td>
<td>314</td>
</tr>
<tr>
<td>9.2.1</td>
<td>Why analyze protein sequences?</td>
<td>314</td>
</tr>
<tr>
<td>9.2.2</td>
<td>The genetic code and codon bias</td>
<td>315</td>
</tr>
<tr>
<td>9.2.3</td>
<td>Look-back time</td>
<td>317</td>
</tr>
<tr>
<td>9.2.4</td>
<td>Nature of sequence divergence in proteins (the PAM unit)</td>
<td>319</td>
</tr>
<tr>
<td>9.2.5</td>
<td>Introns and non-coding DNA</td>
<td>321</td>
</tr>
<tr>
<td>9.2.6</td>
<td>Choosing DNA or protein?</td>
<td>322</td>
</tr>
<tr>
<td>9.3</td>
<td>Construction of phylogenetic trees</td>
<td>323</td>
</tr>
<tr>
<td>9.3.1</td>
<td>Preparation of the data set</td>
<td>323</td>
</tr>
<tr>
<td>9.3.2</td>
<td>Tree-building</td>
<td>329</td>
</tr>
</tbody>
</table>
Contents

Practice 332
Fred R. Oppegoodes 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 PROTEST 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
11.1 Introduction 362
11.2 The relative rate test 364
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
12.10 Conclusions 403
Contents

Section V: Molecular adaptation

13 Natural selection and adaptation of molecular sequences
Oliver G. Pybus and Beth Shapiro

13.1 Basic concepts
13.2 The molecular footprint of selection
  13.2.1 Summary statistic methods
  13.2.2 $d_N/d_S$ methods
  13.2.3 Codon volatility
13.3 Conclusion

14 Estimating selection pressures on alignments of coding sequences
Sergei L. Kosakovsky Pond, Art F. Y. Poon, and Simon D. W. Frost

Theory

14.1 Introduction
14.2 Prerequisites
14.3 Codon substitution models
14.4 Simulated data: how and why?
14.5 Statistical estimation procedures
  14.5.1 Distance-based approaches
  14.5.2 Maximum likelihood approaches
  14.5.3 Estimating $d_S$ and $d_N$
  14.5.4 Correcting for nucleotide substitution biases
  14.5.5 Bayesian approaches
14.6 Estimating branch-by-branch variation in rates
  14.6.1 Local vs. global model
  14.6.2 Specifying branches a priori
  14.6.3 Data-driven branch selection
14.7 Estimating site-by-site variation in rates
  14.7.1 Random effects likelihood (REL)
  14.7.2 Fixed effects likelihood (FEL)
  14.7.3 Counting methods
  14.7.4 Which method to use?
  14.7.5 The importance of synonymous rate variation
14.8 Comparing rates at a site in different branches
14.9 Discussion and further directions

Practice

14.10 Software for estimating selection
  14.10.1 PAML
  14.10.2 ADAPTSITE
## Contents

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  
14.14.1 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 $\beta/\alpha$ 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  
14.16.1 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** 493

Philippe Lemey and David Posada

15.1 Introduction 493  
15.2 Mechanisms of recombination 493
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
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
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

17 The coalescent: population genetic inference using genealogies 551

Allen Rodrigo

17.1 Introduction 551

17.2 The Kingman coalescent 552

17.3 Effective population size 554
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
18.14 Viewing the annotated tree 589
18.15 Conclusion and resources 590
19 LAMARC: Estimating population genetic parameters from molecular data

Theory
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
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

20 Assessing substitution saturation with DAMBE

Theory
Xuhua Xia

20.1 The problem of substitution saturation 615
20.2 Steel’s method: potential problem, limitation, and implementation in DAMBE 616
xviii  Contents

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

Glossary  654
References  672
Index  709
Contributors

Guy Bottu
Belgian EMBnet Node
Brussels, Belgium

Alexei Drummond
Department of Computer Science
University of Auckland
Private Bag 92019
Auckland, New Zealand

Simon Frost
Antiviral Research Center
University of California
150 W Washington St, Ste 100
San Diego, CA 92103, USA

Des Higgins
Conway Institute
University College Dublin
Ireland

Katharina T. Huber
School of Computing Sciences
University of East Anglia
Norwich, UK

John P. Huelsenbeck
Department of Integrative Biology
University of California at Berkeley
3060 Valley Life Sciences Bldg
Berkeley, CA 94720-3140, USA

Sergei Kosakovsky Pond
Antiviral Research Center
University of California
150 W Washington St, Ste 100
San Diego, CA 92103, USA

Mary Kuhner
Department of Genome Sciences
University of Washington
Seattle (WA), USA

Philippe Lemey
Rega Institute for Medical Research
Katholieke Universiteit Leuven
Leuven, Belgium

Darren Martin
Institute of Infectious Disease and Molecular Medicine
Faculty of Health Sciences
University of Cape Town
Observatory 7925
South Africa

Vincent Moulton
School of Computing Sciences
University of East Anglia
Norwich, UK

Fred R. Opperdoes
C. de Duve Institute of Cellular Pathology
Universite Catholique de Louvain
Brussels, Belgium
List of contributors

Art Poon
Antiviral Research Center
University of California
150 W Washington St, Ste 100
San Diego, CA 92103, USA

David Posada
Department of Biochemistry
Genetics and Immunology
University of Vigo
Spain

Oliver Pybus
Department of Zoology
University of Oxford
South Parks Road
Oxford OX1 3PS, UK

Andrew Rambaut
Institute of Evolutionary Biology
University of Edinburgh
Ashworth Laboratories
Kings Building
West Mains Road
Edinburgh EH3 9JT, UK

Allen Rodrigo
School of Biological Sciences
University of Auckland
New Zealand

Fredrik Ronquist
Department of Entomology
Swedish Museum of Natural History
Box 50007, SE-104 05 Stockholm
Sweden

Marco Salemi
Department of Pathology, Immunology, and Laboratory Medicine
University of Florida
Gainesville, Florida
USA

Mika Salminen
HIV Laboratory
National Public Health Institute
Department of Infectious Disease Epidemiology
Helsinki, Finland

Heiko Schmidt
Center for Integrative Bioinformatics Vienna (CIBIV)
Max F. Perutz Laboratories (MFPL)
Dr. Bohr Gasse 9
A-1030 Wien, Austria

Beth Shapiro
Department of Biology
The Pennsylvania State University
326 Mueller Lab
University Park, PA 16802
USA

Korbinian Strimmer
Institute for Medical Informatics Statistics and Epidemiology (IMISE)
University of Leipzig
Germany

Jack Sullivan
Department of Biological Science
University of Idaho
Idaho, USA

David L. Swofford
School of Computational Science and Information Technology
and
Department of Biological Science
Florida State University
Florida, USA

Anne-Mieke Vandamme
Rega Institute for Medical Research
Katholieke Universiteit Leuven
Leuven, Belgium
List of contributors

Yves Van de Peer
VIB / Ghent University
Bioinformatics & Evolutionary Genomics
Technologiepark 927
B-9052 Gent, Belgium

Paul van der Mark
School of Computational Science
Florida State University
Tallahassee, FL 32306-4120, USA

Marc Van Ranst
Rega Institute for Medical Research
Katholieke Universiteit Leuven
Leuven, Belgium

Arndt von Haeseler
Center for Integrative Bioinformatics
Vienna (CIBIV)
Max F. Perutz Laboratories (MFPL)
Dr. Bohr Gasse 9
A-1030 Wien, Austria

Xuhua Xia
Biology Department
University of Ottawa
Ottawa, Ontario
Canada
Foreword

“It looked insanely complicated, and this was one of the reasons why the snug plastic cover it fitted into had the words DON’T PANIC printed on it in large friendly letters.”

Douglas Adams

The Hitch Hiker’s Guide to the Galaxy

As of February 2008 there were 85 759 586 764 bases in 82 853 685 sequences stored in GenBank (Nucleic Acids Research, Database issue, January 2008). Under any criteria, this is a staggering amount of data. Although these sequences come from a myriad of organisms, from viruses to humans, and include genes with a diverse arrange of functions, it can all, at least in principle, be studied from an evolutionary perspective. But how? If ever there was an invitation panic, it is this. Enter The Phylogenetic Handbook, an invaluable guide to the phylogenetic universe.

The first edition of The Phylogenetic Handbook was published in 2003 and represented something of a landmark in evolutionary biology, as it was the first accessible, hands-on instruction manual for molecular phylogenetics, yet with a healthy dose of theory. Up until this point, the evolutionary analysis of gene sequence was often considered something of a black art. The Phylogenetic Handbook made it accessible to anyone with a desktop computer.

The new edition The Phylogenetic Handbook moves the field along nicely and has a number of important intellectual and structural changes from the earlier edition. Such a revision is necessary to track the major changes in this rapidly evolving field, in terms of both the new theory and new methodologies available for the computational analysis of gene sequence evolution. The result is a fine balance between theory and practice. As with the First Edition, the chapters take us from the basic, but fundamental, tasks of database searching and sequence alignment, to the complexity of the coalescent. Similarly, all the chapters are written by acknowledged experts in the field, who work at the coal-face of developing new methods and using them to address fundamental biological questions. Most of the authors are also remarkably young, highlighting the dynamic nature of this discipline.
The biggest alteration from the First Edition is the restructuring into a series of sections, complete with both theory and practice chapters, with each designed to take the uninitiated through all the steps of evolutionary bioinformatics. There are also more chapters on a greater range of topics, so the new edition is satisfyingly comprehensive. Indeed, it almost stands alone as a textbook in modern population genetics. It is also pleasing to see a much stronger focus on hypothesis testing, which is a key aspect of modern phylogenetic analysis. Another welcome change is the inclusion of chapters describing Bayesian methods for both phylogenetic inference and revealing population dynamics, which fills a major gap in the literature, and highlights the current popularity of this form of statistical inference.

The Phylogenetic Handbook will calm the nerves of anyone charged with undertaking an evolutionary analysis of gene sequence data. My only suggestion for an improvement to the third edition are the words DON’T PANIC on the cover.

Edward C. Holmes
June 12, 2008
Preface

The idea for *The Phylogenetic Handbook* was conceived during an early edition of the Workshop on Virus Evolution and Molecular Epidemiology. The rationale was simple: to collect the information being taught in the workshop and turn it into a comprehensive, yet simply written textbook with a strong practical component. Marco and Annemie took up this challenge, and, with the help of many experts in the field, successfully produced the First Edition in 2003. The resulting text was an excellent primer for anyone taking their first computational steps into evolutionary biology, and, on a personal note, inspired me to try out many of the techniques introduced by the book in my own research. It was therefore a great pleasure to join in the collaboration for the Second Edition of *The Phylogenetic Handbook*.

Computational molecular biology is a fast-evolving field in which new techniques are constantly emerging. A book with a strong focus on the software side of phylogenetics will therefore rapidly grow a need for updating. In this Second Edition, we hope to have satisfied this need to a large extent. We also took the opportunity to provide a structure that groups different types of sequence analyses according to the evolutionary hypothesis they focus on. Evolutionary biology has matured into a fully quantitative discipline, with phylogenies themselves having evolved from classification tools to central models in quantifying underlying evolutionary and population genetic processes. Inspired by this, the Second Edition provides a broader coverage of techniques for testing models and trees, detecting recombination, the analysis of selective pressure and genealogy-based population genetics. Changing the subtitle to *A Practical Approach to Phylogenetic Inference and Hypothesis Testing* emphasizes this shift in focus. Thanks to novel contributions, we also hope to have addressed the need for a Bayesian treatment of phylogenetic inference, which started to gain a great deal of popularity at the time the content for the First Edition was already fixed.

Following the philosophy of the First Edition, the book includes many step-by-step software tutorials using example data sets. We have not used the same data sets throughout the complete Second Edition; not only is it difficult to find data sets that
Preface

consistently meet the assumptions or reveal interesting aspects of all the methods described, but we also feel that being confronted with different data with their own characteristics adds educational value. These data sets can be retrieved from www.thephylogenetichandbook.org, where other useful links listed in the book can also be found. Furthermore, a glossary has been compiled with important terms that are indicated in italics and boldface throughout the book.

We are very grateful to the researchers who took the time to contribute to this edition, either by updating a chapter or writing a novel contribution. I hope that my persistent pestering has not affected any of these friendships. We would like to thank Eddie Holmes in particular for writing the Foreword to the book. It has been a pleasure to work with Katrina Halliday and Alison Evans of Cambridge University Press. We also wish to thank those who supported our research and the work on this book: the Flemish “Fonds voor Wetenschappelijk Onderzoek”, EMBO and Marie Curie funding. Finally, we would like to express our thanks to colleagues, family and friends onto whom we undoubtedly projected some of the pressure in completing this book.

Philippe Lemey