The Phylogenetic Handbook

A Practical Approach to DNA and Protein Phylogeny

Edited by

Marco Salemi

University of California, Irvine
(Formerly of Katholieke Universiteit Leuven)

and

Anne-Mieke Vandamme

Rega Institute for Medical Research, Katholieke Universiteit Leuven, Belgium

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011-4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

First published 2003

Printed in the United Kingdom at the University Press, Cambridge

Typefaces Minion 10.5/14 pt. and Formata *System* LaTeX 2$_\varepsilon$ [TB]

A catalog record for this book is available from the British Library.

Library of Congress Cataloging in Publication Data

The phylogenetic handbook : a practical approach to DNA and protein phylogeny /
edited by Marco Salemi, Anne-Mieke Vandamme.

p. cm.

Includes bibliographical references and index.

ISBN 0-521-80390-X

1. DNA – Analysis – Handbooks, manuals, etc. 2. Proteins – Analysis – Handbooks, manuals, etc. 3. Cladistic analysis – Handbooks, manuals, etc. I. Salemi, Marco, 1968–II. Vandamme, Anne-Mieke, 1960–

QP624 .P485 2003
572.8′633 – dc21 2002073927

ISBN 0 521 80390 X hardback

Contents

Foreword

Theodosius Dobzhansky (1973) wisely said, "Nothing in biology makes sense except in the light of Evolution." This truism is so often repeated that it is nearly a mantra and, with the complete genomes of many organisms being completed nearly daily, all kinds of people, but especially molecular biologists and informaticists, are rediscovering that truth. And with that discovery they are coming to need to know the tools of the trade that have been under development for nearly forty years. This book is for them in particular but it has much that, except for polymaths, may be useful even to the cognoscenti.

The book has grown out of Drs. Vandamme's and Salemi's annual course in these methods at the Katholieke Universiteit in Leuven, Belgium, where they have produced an exceptional workshop for eight years that does for Europe what a similar outstanding and long-running workshop at Woods Hole did for the United States and Canada. But the latter has not created a book like this.

The coverage is comprehensive. Topics touched upon include databases, multiple alignments, nucleotide substitution models, phylogeny inference methods (such as distance, maximum likelihood, and maximum parsimony), post-phylogenetic information (such as molecular clocks and selection), and useful subsidiary statistical techniques (such as bootstrapping and likelihood ratio tests).

Each of the major sections is written by an expert in the field, and each such section is divided into two major subsections, theory and practice. This permits the novice to proceed with his analysis without having to master the theory. That is, of course, very dangerous in this field where so many methods have different assumptions and the failure of any one of those assumptions (clocklike behavior, all sites equally mutable, all substitutions neutral) can reduce your analysis to rubbish, if untrue, which they frequently are. Still, there are people like that and we may hope that a good text such as this, with its many caveats and generally simple prose, will reduce the published trash.

The material is enhanced by the use of specific examples from which you can see what to expect, and see if you can get the same answer, and then try your own data

to see if anything strange has happened. The examples also aid in locating what you need to find in the text.

Another aspect of the book that enhances its utility for the reader is the repeated use of the same three data sets, even by different authors, to illustrate the methods. This increases immensely the value of the exercises. This is especially true when the results from different methods are ostensibly for the same desired end, and one gets to see how they differ and why (or at least to worry about it).

The example data sets used in the book can be downloaded from the book's website [http://www.kuleuven.ac.be/aidslab/phylogenybook.htm]. On the website the reader can also find useful links to the major phylogeny resources on the internet, as well as the results of all the analyses discussed in the text, including phylogenetic trees, unaligned and aligned sequences, and so forth.

It is appropriate to compare this work with others in the general area. The first two are by Weir (1990) and by Waterman (1995). They are both highly theoretical and quite capable of turning off many biologists quickly (although Waterman's book can be highly engaging as in his recounting of the efforts of George Gamow to predict that the genetic code was a commaless code). At the other extreme, Hall (2001) is really simple-minded enough (intentionally so) that a bright senior could easily master the methods. However, the Hall book lacks the comprehensiveness of the Salemi and Vandamme work. Two other good books, Li (1998), and Page and Holmes (1998), are largely theoretical although they make a great effort to make the subject palatable to the biologist who is mathematically challenged. In sum, there is no other book even trying to occupy the niche of this one.

In conclusion, this is a relatively easy-to-use workbook for phylogenetics, especially if the index is properly looked to (I haven't seen it). However, I have to present a strongly worded negative comment. Although tables and figures in the book have titles, many have no legends and many of the remainder have poor legends. For example, numbers normally have dimensions, (such as nucleotide differences per hundred nucleotide positions), that should have been given. Figures and tables should be as self-sufficient as is reasonable. This is not true here. Let us hope this is corrected in the next printing, which I am sure this book will achieve.

Walter M. Fitch
December 27, 2002

REFERENCES

Dobzhansky, Theodosius (1973). Nothing in biology makes sense except in the light of evolution. *American Biology Teacher*, 125–129.

Weir, Bruce S. (1990). *Genetic Data Analysis*, Sunderland, MA: Sinauer Associates.

Waterman, Michael S. (1995). *Introduction to Computational Biolog: Maps, sequences and genomes.* New York: Chapman and Hall.

Hall, Barry G. (2001). *Phylogenetic Trees Made Easy*, Sunderland, MA: Sinauer Associates.

Li, Wen-Hsiung (1997). *Molecular Evolution*, Sunderland, MA: Sinauer Associates.

Page, Roderic D. M. and Edward C., Holmes (1998). *Molecular Evolution: A Phylogenetic Approach.* London: Blackwell Science.

Note: During the writing of this book the alpha release of the new version of PHYLIP, PHYLIP 3.6 has been made available on the PHYLIP web page. All the exercises with PHYLIP refer to version 3.5, but additional exercises covering PHYLIP v3.6 can be found on the book website.

Contributors

Guy Bottu
Belgian EMBnet Node
Brussels, Belgium

Walter Fitch (Foreword)
Ecology and Evolutionary Biology
University of California, Irvine
California, USA

Takashi Gojobori
Center for Information Biology
and DNA Data Bank of Japan
National Institute of Genetics
Mishima, Japan

Arndt von Haeseler
Heinrich-Heine-Universitat Düsseldorf
Institut für Bioinformatik
Düsseldorf, Germany

Des Higgins
Department of Biochemistry
University College
Cork, Ireland

Mary K. Kuhner
Department of Genome Sciences
University of Washington
Washington, USA

Vincent Moulton
Physics and Mathematics Department
Mid Sweden University
Sundsvall, Sweden

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

Yves Van de Peer
Department of Plant Systems Biology
Flanders Interuniversity Institute for
Biotechnology (VIB)
Ghent University
Ghent, Belgium

David Posada
Departamento de Bioquímica, Xenética e
Inmunoloxía
Facultade de Ciencias
Universidade de Vigo
Vigo, Spain

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

Marco Salemi
Rega Institute for Medical Research
Katholieke Universiteit Leuven
Leuven, Belgium
and
Ecology and Evolutionary Biology
University of California, Irvine
California, USA

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

Korbinian Strimmer
Department of Statistics
University of München
München, Germany

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

Yoshiyuki Suzuki
Center for Information Biology
and DNA Data Bank of Japan
National Institute of Genetics
Mishima, Japan

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

Xuhua Xia
Biology Department
University of Ottawa
Ottawa, Ontario
Canada

Zheng Xie
Institute of Environmental Protection
Hunan University
China

5

Phylogeny inference based on distance methods

THEORY

Yves Van de Peer

5.1 Introduction

In addition to ***maximum parsimony*** (***MP***) and likelihood methods (see Chapters 6 and 7), pairwise ***distance methods*** form the third large group of methods to infer evolutionary trees from sequence data (Figure 5.1). In principle, distance methods try to fit a tree to a matrix of pairwise ***genetic distances*** (Felsenstein, 1988). For every two sequences, the distance is a single value based on the fraction of positions in which the two sequences differ, defined as ***p-distance*** (see Chapter 4). The p-distance is an underestimation of the true genetic distance because some of the aligned nucleotides are the result of multiple events. Indeed, because mutations are fixed in the genes, there has been an increasing chance of multiple substitutions occurring during evolution at the same sequence position. Therefore, in distance-based methods, one tries to estimate the number of substitutions that have actually occurred by applying a specific ***evolutionary model*** that makes assumptions about the nature of evolutionary changes (see Chapter 4). When all the pairwise distances have been computed for a set of sequences, a tree topology can then be inferred by a variety of methods (Figure 5.2).

Correct estimation of the genetic distance is crucial and, in most cases, more important than the choice of method to infer the tree topology. Using an unrealistic evolutionary model can cause serious artifacts in tree topology, as previously shown in numerous studies (e.g., Olsen, 1987; Lockhart et al., 1994; Van de Peer et al., 1996; see also Chapter 10). However, because the exact historical record of events that occurred in the evolution of sequences is not known, the best method for estimating the genetic distance is not necesserily self-evident (see Chapter 9).

	Character-based methods	Noncharacter-based methods
Methods based on an explicit model of evolution	Maximum-likelihood methods	Pairwise-distance methods
Methods not based on an explicit model of evolution	Maximum-parsimony methods	

Figure 5.1 Pairwise distance methods are non-character-based methods that make use of an explicit substitution model.

Substitution models are discussed in Chapters 4, 9, and 10. Chapters 7 and 10 discuss how to select the best-fitting evolutionary model for a given data set of nucleotide or amino-acid aligned sequences in order to get an accurate estimation of the genetic distances. In the following sections, it is assumed that genetic distances were estimated using an appropriate evolutionary model, and some of the methods used for inferring tree topologies on the basis of these distances are briefly outlined. However, by no means should this be considered a complete discussion of distance methods; additional discussions are in Felsenstein (1982), Swofford et al. (1996), Li (1997), and Page and Holmes (1998).

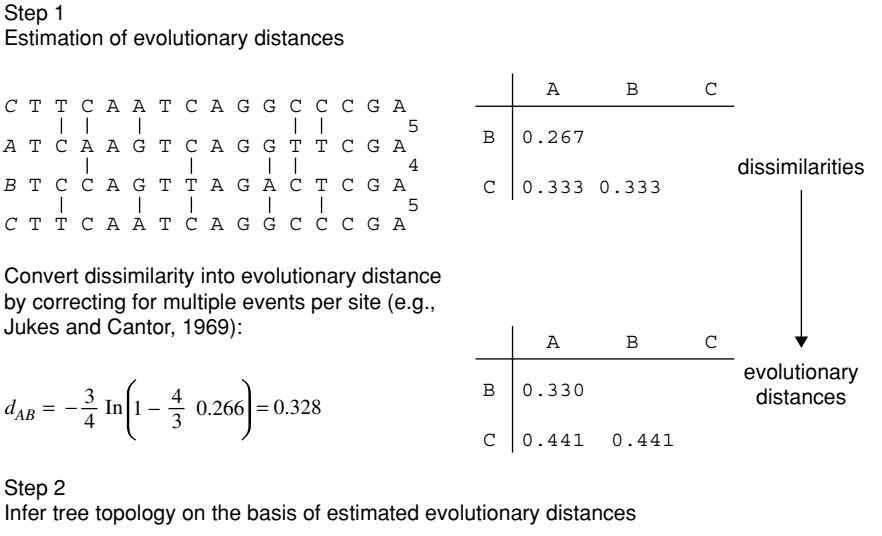


Figure 5.2 Distance methods proceed in two steps. First, the evolutionary distance is computed for every sequence pair. Usually, this information is stored in a matrix of pairwise distances. Second, a tree topology is inferred on the basis of the specific relationships between the distance values.

5.2 Tree-inferring methods based on genetic distances

The main distance-based tree-building methods are cluster analysis and minimum evolution. Both rely on a different set of assumptions, and their success or failure in retrieving the correct phylogenetic tree depends on how well any particular data set meets such assumptions.

5.2.1 Cluster analysis (UPGMA and WPGMA)

Clustering methods are tree-building methods that were originally developed to construct taxonomic phenograms (Sokal and Michener, 1958; Sneath and Sokal, 1973); that is, trees based on overall phenotypic similarity. Later, these methods were applied to phylogenetics to construct ***ultrametric trees***. ***Ultrametricity*** is satisfied when, for any three ***taxa***, A, B, and C,

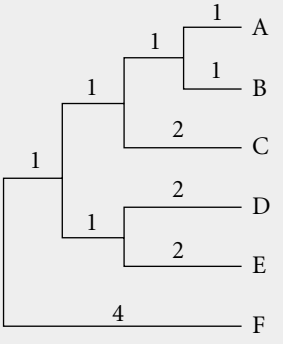
$$d_{AC} \leq \max(d_{AB}, d_{BC}). \tag{5.1}$$

In practice, Equation 5.1 is satisfied when two of the three distances under consideration are equal and as large (or larger) as the third one. Ultrametric trees are ***rooted trees*** in which all the end nodes are equidistant from the root of the tree, which is only possible by assuming a ***molecular clock*** (see Chapters 1 and 10). Clustering methods such as the ***unweighted-pair group method with arithmetic means*** (***UPGMA***) or the ***weighted-pair group method with arithmetic means*** (***WPGMA***) use a sequential clustering algorithm. A tree is built in a stepwise manner, by grouping sequences or groups of sequences – usually referred to as ***operational taxonomic units*** (***OTUs***) – that are most similar to each other; that is, for which the genetic distance is the smallest. When two OTUs are grouped, they are treated as a new single OTU (Box 5.1). From the new group of OTUs, the pair for which the similarity is highest is again identified, and so on, until only two OTUs are left. The method applied in Box 5.1 is actually the WPGMA, in which the averaging of the distances is not based on the total number of OTUs in the respective clusters. For example, when OTUs A, B (which have been grouped before), and C are grouped into a new node '*u*', then the distance from node '*u*' to any other node '*k*' (e.g., grouping D and E) is computed as follows:

$$d_{uk} = \frac{d_{(A,B)k} + d_{Ck}}{2} \tag{5.2}$$

Conversely, in UPGMA, the averaging of the distances is based on the number of OTUs in the different clusters; therefore, the distance between '*u*' and '*k*' is

Box 5.1 Cluster analysis (Sneath and Sokal, 1973)

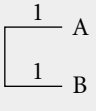


	A	B	C	D	E
B	(2)				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

N = 6

Cluster analysis proceeds as follows:

1. Group together (cluster) these OTUs for which the distance is minimal; e.g., A and B together. The depth of the divergence is the distance between A and B divided by 2.



2. Compute the distance from cluster (A, B) to every other OTU.

$$d_{(AB)C} = (d_{AC} + d_{BC})/2 = 4$$
$$d_{(AB)D} = (d_{AD} + d_{BD})/2 = 6$$
$$d_{(AB)E} = (d_{AE} + d_{BE})/2 = 6$$
$$d_{(AB)F} = (d_{AF} + d_{BF})/2 = 8$$

	(AB)	C	D	E
C	4			
D	6	6		
E	6	6	(4)	
F	8	8	8	8

Repeat Steps 1 and 2 until all OTUs are clustered (repeat until N = 2).

N = N − 1 = 5

1. Group together (cluster) these OTUs for which the distance is minimal; e.g., group D and E together. Alternatively, (AB) could be grouped with C.